Protein Engineering, Design & Selection vol. 20 no. 6 pp. 285–299, 2007
Published online June 30, 2007 doi:10.1093/protein/gzm021

Detecting similarities among distant homologous proteins by comparison of domain flexibilities

Alessandro Pandini[1], Giancarlo Mauri[2], Annalisa Bordogna[1] and Laura Bonati[1,3]

[1]Dipartimento di Scienze dell'Ambiente e del Territorio and [2]Dipartimento di Informatica Sistemistica e Comunicazione, Università degli Studi di Milano-Bicocca, 20126 Milano, Italy

[3]To whom correspondence should be addressed.
E-mail: laura.bonati@unimib.it

Aim of this work is to assess the informativeness of protein dynamics in the detection of similarities among distant homologous proteins. To this end, an approach to perform large-scale comparisons of protein domain flexibilities is proposed. CONCOORD is confirmed as a reliable method for fast conformational sampling. The root mean square fluctuation of alpha carbon positions in the essential dynamics subspace is employed as a measure of local flexibility and a synthetic index of similarity is presented. The dynamics of a large collection of protein domains from ASTRAL/SCOP40 is analyzed and the possibility to identify relationships, at both the family and the superfamily levels, on the basis of the dynamical features is discussed. The obtained picture is in agreement with the SCOP classification, and furthermore suggests the presence of a distinguishable familiar trend in the flexibility profiles. The results support the complementarity of the dynamical and the structural information, suggesting that information from dynamics analysis can arise from functional similarities, often partially hidden by a static comparison. On the basis of this first test, flexibility annotation can be expected to help in automatically detecting functional similarities otherwise unrecoverable.

Keywords: ASTRAL/SCOP/CONCOORD/Essential Dynamics/domain flexibility/molecular simulations

Introduction

A central problem in protein studies is inferring functional analogies from sequence or structure similarities. At a sequence level, it was demonstrated that the goal of detecting similar among non-similar structures and inferring homology between proteins with more than 35–40% of residue identity is achievable without additional information. Below 30% of the residue identity for an average 100 residue proteins, there is a rapid transition to a more difficult problem and a sudden explosion of false positives in homology detection (Rost, 1999). When tertiary structures are available, the employment of structural alignment, supported by accurate statistical estimates, allows to detect similarities and derive high-quality sequence alignments. Anyway the accuracy of the most sensitive methods is comparable to reliable sequence-based methods, therefore highlighting a similar tendency in reporting false-positive for difficult structural comparisons (Pearson and Sierk, 2005). This evidence suggests the need to employ additional information to improve protein comparison.

Structural flexibility is essential for most of the proteins to perform their biological activity (Gerstein *et al.*, 1994) and it often highlights functional specificities. Therefore, the additional source of information from the protein intrinsic flexibility looks promising to detect similarities.

Experimental information on flexibility in the form of structural fluctuations can be retrieved by nuclear magnetic resonance (NMR) and X-ray diffraction techniques. The former has been increasingly employed, but the range of structures that have been determined is still relatively small; additionally, the physical interpretation of the experimental data is often difficult. The most widely employed experimental measure of atomic fluctuations comes indeed from X-ray crystallographic B-factors (Frauenfelder *et al.*, 1979). These describe the isotropic mean square displacement of the atom from its average position. Therefore, the represented flexibility includes contributions from internal protein motions, from translation and rotation of the whole molecule in the unit cell, from relative translation of the unit cell itself; additionally, lattice distortion and refinement errors can affect the values. Consequently, the employment of B-factors as a direct measure of intrinsic molecular flexibility can be problematic. The presence of different sources of motion, aside from the internal flexibility, suggests carefulness in the functional interpretation. Moreover, by comparing computational simulation of a single molecule in explicit solvent with that of a whole unit cell, recent studies demonstrated that crystal packing is greatly affecting the flexibility (Eastman *et al.*, 1999; Meinhold and Smith, 2005). This effect is stronger for exposed loops and less structured regions, which are often involved in functional activity. Very recently, it was also demonstrated that B-factor flexibility is highly correlated with local protein packing density, suggesting that the informativeness of this index is comparable to that of the mean atomic coordinates (Halle, 2002).

Another avenue to derive protein flexibility is employing molecular simulations, especially molecular dynamics (MD). Explicit solvent simulations can generate thermodynamic ensembles of structures representing the different states of the native protein. Flexibility calculated from these ensembles is not affected by crystal packing bias and is supposed to properly describe solvent exposed region (Meinhold and Smith, 2005). In practice, the descriptive power of the simulations is limited by the timescale: several studies addressed the problem of convergence of atomic fluctuations, highlighting how nanoseconds simulations are suitable for description of structured region, but insufficient to extensively sample loop conformations (Hunenberger *et al.*, 1995; Eastman *et al.*, 1999; Meinhold and Smith, 2005). Convergence on correlated motions for non-local atoms would require extensive sampling

up to microseconds scale (Meinhold and Smith, 2005). To address this problem, a major interest has arisen for fast conformational sampling methods. These approaches try to obtain reliable sampling of native ensembles within an affordable computational time and indeed are suitable to derive simulated flexibilities with acceptable accuracy (Tai, 2004).

Simulated flexibility has already been employed to investigate the functionality of single proteins of interest, their mutants or small collections of strictly related functional analogous. These studies are mainly directed to highlight the conservation of dynamical features across a fold (Grottesi and Sansom, 2003) or a superfamily (Vreede *et al.*, 2003), or suggest a more complex combination of conservation and specialization at the superfamily level (Pandini and Bonati, 2005). Recently, a promising approach to large-scale investigation of flexibility has been proposed, which employs the Gaussian network model (GNM) (Bahar *et al.*, 1997), instead of simulations. This approach has been extended to comparisons at the superfamily level (Maguid *et al.*, 2005) and also made available as a database resource (Yang *et al.*, 2005).

Moreover, some studies have effectively correlated protein flexibilities, as derived from normal mode analysis of the elastic network model (ENM), to sequence conservation among protein families (Zheng *et al.*, 2005).

Therefore, it was demonstrated that inferring similarities among proteins with the same functionality is often achievable by comparison of their flexibilities. Questions arise if this is a general rule and if simulated flexibilities can be employed effectively for protein comparison, especially in the cases of distant homologous proteins.

In this framework, the aim of this work is to perform an analysis on the informativeness of a measure of protein flexibility derived from molecular simulations.

To this end, a procedure is proposed for the analysis of relationships among distant homologous proteins at the domain level. This is aimed at a large-scale analysis and consequently employs a fast conformational sampling method and a simple and synthetic index of similarity between domain flexibilities, calculated on the residue base.

To assess the reliability of this procedure, the dynamics of a collection of protein domains from ASTRAL/SCOP40 was analyzed and the possibility to identify relationships, at both the family and the superfamily SCOP levels, on the basis of the dynamical features is discussed. The discrimination power of the procedure was also analyzed in relation to that of some structure comparison methods on the same test set.

Materials and methods

Conformational sampling

The data on the local flexibility of each protein were extracted from a collection of structures representing a statistical ensemble in the neighborhood of the starting structure. The ensemble was obtained by CONCOORD (de Groot *et al.*, 1997) runs. This is a computational method to generate conformers satisfying a list of distance constraints derived from a starting structure. The imposed constraints are a good approximation for the upper and lower bound of each interatomic distance and the stochastic sampling of the associated values allows a fast generation of structures. For the calculation of interatomic distances, non-bonded cut-off radius was set to the sum of vdW radii plus 4.00 Å and the minimal number of distances per atom to 100. Secondary structure assignment was made according to Kabsch and Sander (Kabsch and Sander, 1983). Hydrogen atoms were not included in the calculation. The maximum number of iterations for the generation of a structure satisfying all distance constraints was set to 500. For each protein, a collection of 2000 structures was generated.

Some selected domains were also simulated by MD. The trajectories were generated and analyzed by GROMACS 3.2.1 (Berendsen *et al.*, 1995; Lindahl *et al.*, 2001). All structures were inserted in a SPC water (Berendsen *et al.*, 1981) dodecahedral box and simulated with periodic boundary conditions. The dimension of the box was set to allow at least 0.8 nm between the protein and the box faces. Solvent was relaxed with 5 ps MD simulation, during which the protein degrees of freedom were restrained. After neutralizing the systems with the appropriate number of counter ions, a short minimization with steepest descent was performed up to convergence on maximum force lower than 1000 kJ/mol nm. The resulting systems were simulated for 16 ns with the GROMOS96 43a2 version of the GROMOS force field as available in the GROMACS package. Simulations were performed in the NPT ensemble and long-range electrostatic interactions were calculated with the particle mesh Ewald summation method (Darden *et al.*, 1993) to gain a more accurate description. Van der Waals interactions were described by a 6–12 Lennard–Jones potential with distance cut-off at 0.9 nm; neighbor lists were employed with a list cut-off of 0.9 nm and update frequency every 10 steps. A thermal bath was independently coupled with protein and solvent by employment of a Berendsen thermostat at 300 K and a coupling period of 0.1 ps. Internal degrees of freedom of water were constrained by the SHAKE algorithm (Ryckaert *et al.*, 1977), while all bond distances in the protein were constrained by the LINCS algorithm (Hess *et al.*, 1997). To allow a wider time step, the interacting site method (Feenstra *et al.*, 1999) was employed; therefore, it was possible to increase the integration step up to 4 fs and obtain stable simulations. During simulations, configurations and velocities were recorded every 1 ps.

Flexibility analysis and representation

The extraction of the data on the local flexibility of each protein from CONCOORD ensembles and MD simulations was performed after essential dynamics (ED) analysis. ED analysis is a widely applied technique based on principal component analysis (PCA) of conformational ensembles (Amadei *et al.*, 1993). It is aimed to extract informative directions of motion in a multidimensional space and allows to both reduce the overall complexity of the simulation and isolate the important motions with a putative functional meaning. ED application involves: diagonalization of the covariance matrix of the positional fluctuations of atoms; projection of original data on the eigenvectors to generate $3N$ principal components; separation of the simulation space in the more informative 'essential subspace' and in the constrained subspace, by evaluating the amount of variance contained in the first eigenvectors.

Only the Cα atoms were included in the analysis, because it was demonstrated that this reduction of the analysis can retain all the relevant information needed to separate the

essential subspace and identify the important modes in the protein dynamics (Amadei *et al.*, 1993). In the definition of the dimensionality of the essential subspace, two criteria were employed: the fraction of total motion described by the reduced subspace and the distribution of motion along the eigenvectors. The former, computed as the sum of eigenvalues for the included eigenvectors and expressed as percentage of the total fluctuation of the Cα atoms, describes the amount of variance retained by the reduced representation of the system space; the latter is evaluated by projection of motion on the single directions and calculation of the corresponding distributions of motion.

The local flexibility of each protein was then reported as the root mean square fluctuation (RMSF) on the positions of the Cα atoms as calculated from the coordinate of the system in the essential subspace. The RMSF of the Cα atom i is a measure of the deviation between its position, r_i, and its time-averaged position:

$$\mathrm{RMSF}(i) = \sqrt{\langle r_i^2 \rangle - \langle r_i \rangle^2}$$

with $< \ldots >$ indicating a time-average.

For some selected domains, RMSF profiles were also derived from the crystallographic B-factors, B_i, using the relation (Hunenberger *et al.*, 1995):

$$\mathrm{RMSF}(i) = \sqrt{\frac{3}{8\pi^2} B_i}$$

Flexibility comparison

To run the comparison among protein domains, the vectorial representation of the flexibility obtained by the residue-based RMSF was employed. For each pairwise comparison, the two structures were aligned and the structural equivalent positions in the alignment were annotated. Then, the RMSF vectors were filtered to include only these positions. Two programs for structural alignment were employed: the standalone version of the DALI method (Holm and Sander, 1993) called DALILite (Holm and Park, 2000), and Structal (Gerstein and Levitt, 1998).

A median-based method to detect outliers (Iglewicz and Hoaglin, 1993) was used. First, the median of the Cα RMSF, $\tilde{x}$, was determined and then the median of absolute displacements (MAD) from the median was determined. A M_i value for each RMSF was calculated as follows:

$$M_i = 0.6745 \frac{x_i - \tilde{x}}{\mathrm{MAD}}$$

where x_i is the Cα RMSF for the ith residue and multiplication by 0.6745 is used because the expected value of MAD is 0.6745σ for large sample sizes. A M_i value of 3.5 was used to define the outliers to be excluded from the comparison. For the pair of resulting RMSF vectors, the R Pearson correlation coefficient between the two vectors was calculated:

$$r(x,y) = \frac{\sum_i (x_i - < x >)(y_i - < y >)}{\sqrt{\sum_i (x_i - < x >)^2 \sum_i (y_i - < y >)^2}}$$

where $<\cdot\cdot>$ indicates the average value. The resulting value was taken as the index of flexibility correlation (FC) between the pair of domains.

Hierarchical cluster analysis with complete linkage method (Everitt, 1974) was performed on the FC and RMSD matrices. Since FC is a measure of proximity (0 = different flexibility, 1 = same flexibility), FC was converted into a distance index (through the transformation of 1-FC).

Statistical significance

To assess the amount of information in the comparison, the K correlation index (Todeschini, 1997) was calculated. This is an index of correlation among a set of variables. It can be derived from a correlation or covariance matrix of the variables from this relation:

$$K = \frac{\sum_m |\lambda_m / \sum_m \lambda_m| - 1/p}{2(p-1)/p} \times 100$$

where the K index is computed after principal component analysis on the matrix. λ_m are the eigenvalues and p is their number. The index is expressed as percentage of correlation, where 0% indicates complete independence of the variables and 100% the total linear dependence.

In this application, the K correlation was derived from: the FC matrix including all the inter-domain comparisons; a set of FC matrices representative of a random background distribution; a model similarity matrix associated to an hypothetical optimal separation of proteins in families as derived from the SCOP classification.

Analysis of performance

The performance of the FC similarity index was assessed by receiver operating characteristic (ROC) curves (Gribskov and Robinson, 1996) and compared with the performance of a pure geometrical index, the root mean square displacement (RMSD) on equivalent Cα atoms as identified by structural alignment.

The ROC curve is a widely employed statistic to evaluate the ability of a classifier to adhere to a gold standard. According to the gold standard, a group of comparisons are annotated as true (T) or false (F), if they satisfy or not a relationship criterion. The comparisons are then evaluated by the classifier and sorted by a quality measure (score). Given a threshold value for the quality measure, all the comparisons above this value are predicted as positive (showing a relationship). Comparisons positive for the classifier and true for the gold standard are denoted as true positive (TP), while comparisons that are positive but false are labeled false positive (FP). For each classifier, the true positive rate (TPR) and false positive rate (FPR) are calculated for a range of thresholds and the corresponding curve is plotted in a (TPR, FPR) plane. TPR is the ratio of TP on all positive comparisons, while FPR is the ratio of FP on all negative comparisons. For a reliable method, it is desirable to maximize TPR while minimizing FPR. The best method would have a TPR = 1 and FPR = 0 for all thresholds: this translates to the ordinate axis in (TPR, FPR) plane. A concise measure of the accuracy of the classifier is given by the area under the curve (AUC), which is in turn the probability of obtaining a correct classification.

In the assessment of the FC and RMSD indexes, the SCOP classifications for family and superfamily were employed as the gold standards for two independent tests. A cutoff difference of 0.1 in AUC values was employed to define that an index outperforms the other.

Data processing and visualization

Graphs were generated using R 2.1.0 (R Development Core Team, 2003) and MATLAB 7.1.0. Cluster analysis was performed with MATLAB Statistical Toolbox 5.1. The analysis of performance was carried out by employment of the ROCR package (Sing *et al.*, 2005). The molecular model image was generated using PyMOL (DeLano, 2002).

Results

Test set selection

To test an approach for large-scale analysis of protein flexibility, a test case that is representative of the structural information available at this time in the on-line databases was selected. Protein domains were extracted from the SCOP database (Murzin *et al.*, 1995) and, in particular, from the ASTRAL/SCOP40 compendium (release 1.67), (Brenner *et al.*, 2000; Chandonia *et al.*, 2002, 2004) which offers a non-redundant set of protein domains with $<40\%$ sequence similarity. This gave the possibility to test the procedure on distant homologous proteins.

In the SCOP hierarchy, *families* contain protein domains that share a clear common evolutionary origin, as evidenced by sequence identity ($\geq 30\%$) or extremely similar structure and function, *superfamilies* consist of families whose proteins share very common structure and function, and therefore there is reason to believe that these are evolutionary related, *folds* consist of one or more superfamilies that share a common core structure, and finally, depending on the type and organization of secondary structural elements, folds are grouped in the four major *classes:* all α, all β, α/β and $\alpha + \beta$.

A group of 215 protein domains belonging to 8 folds in the $\alpha + \beta$ class were employed as the test set (Table I). Each of the selected folds satisfied the following requirements: containing a number of proteins ranging from 10 to 50 and more than five families; possessing an average proteins/families ratio higher than 2. In particular, the d.16 fold ('*FAD-linked reductases, C-terminal domain*') was used to discuss in more details some of the key methodological choices, whereas the entire set was employed for a global evaluation of the performance of the proposed procedure for flexibility comparison. The choice of the d.16 fold was motivated by the extensive literature on the role of flexibility in the biological activity of its proteins. The PDB ID of the d.16 domains and the ranges of residues included in the study are reported in Table II. In this case, as the fold includes 18 domains in 6 families belonging to a unique superfamily, only the relationships among the flexibilities of homologous domains belonging to the same family (intrafamily) or to different families in the same superfamily (interfamily) can be analyzed. The 18 proteins account for 153 pairwise comparisons with 24 intrafamily (15.7%) and 129 interfamily (84.3%) pairs.

The domains belonging to this fold are generally located at the C-terminus of flavoproteins devoted to redox reactions of small substrates (Fraaije and Mattevi, 2000; Miura, 2001). The fold is characterized, at one side, by a β-sheet of 4–5 strands that faces the active site and the FAD cofactor and, on the opposite side, by a long α-helix and a varying number of small α-helices (Fig. 1). This latest part is the most

Table I. Test set from ASTRAL/SCOP40

Fold	SCOP ID	No. of superfamilies	No. of familes	No. of proteins
Cysteine proteinases	d.3	1	10	22
Ribosomal protein S5 domain two-like	d.14	1	11	30
FAD linked reductases C-terminal domain	d.16	1	6	18
Bacillus chorismate mutase-like	d.79	7	11	25
ATP-grasp	d.142	2	11	23
Protein kinase-like (PK-like)	d.144	1	6	35
Ntn hydrolase-like	d.153	2	6	33
C-type lectin-like	d.169	1	6	29
Total number		16	77	215

Table II. The FAD-linked reductases C-terminal domain fold (d.16)

Family	SCOP ID	PDB ID	Chain:residues
GMC oxidoreductases	d.16.1.1	*1GPE*	*A:329–524*
		1N4W	A:319-450
		1JU2	A:294-463
		1KDG	A:513-693
PHBH-like	d.16.1.2	*1K0I*	*A:174–275*
		1PN0	A:241-341
D-aminoacid oxidase-like	d.16.1.3	*1C0P*	*A:1194–1288*
		1VE9	A:195–287
		1EL5	A:218–321
		1NG4	A:219–306
L-aminoacid/polyamine oxidase	d.16.1.5	*1S3E*	*A:290–401*
		1B37	A:294–405
		1F8R	A:320–432
		1PJ5	A:220–338
		1SEZ	A:330–441
GDI-like	d.16.1.6	*1D5T*	*A:292–388*
		1LTX	R:445–557
UDP-galactopyranose mutases	d.16.1.7	*1I8T*	*A:245–313*

Representative domains for method assessment are in italics.

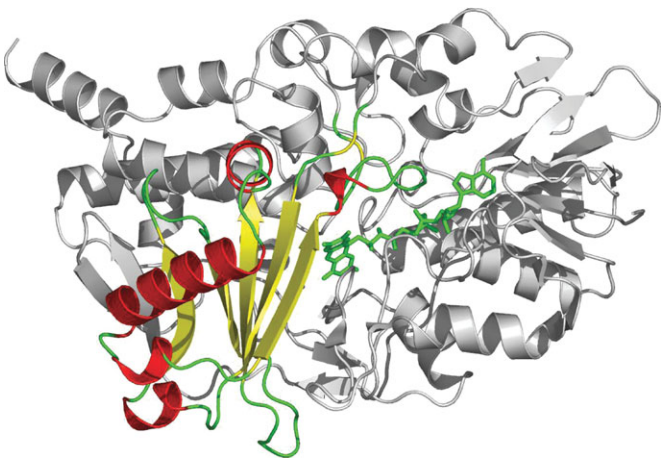

Fig. 1. Cartoon representation of the monoamine oxidase B (PDB ID: 1SE3), taken as a reference for showing the d.16 fold characteristics. The region encompassing the d.16 fold (residue A:290–401) is colored according to the secondary structure attribution: α-helices in red and β-strands in yellow. The flavin-adenine dinucleotide is rendered in green sticks.

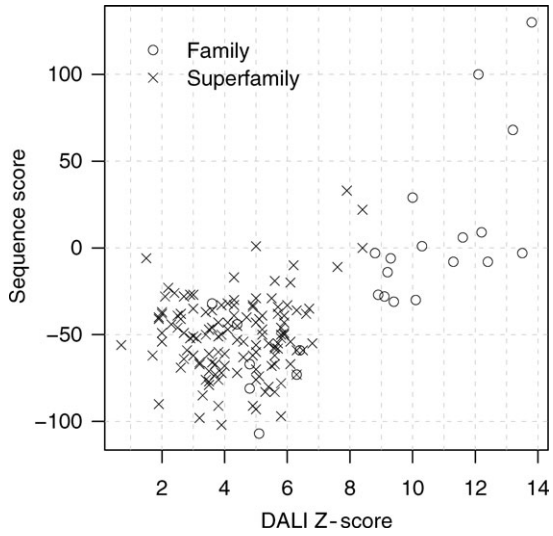

Fig. 2. Extent of sequence and structure similarities among the domains in the ASTRAL/SCOP d.16 fold. For each pairwise comparison, the raw sum of the BLOSUM62 scores for the structurally equivalent positions is plotted against the value of DALI *Z*-score. Comparisons between pairs of proteins belonging to the same family, as reported in SCOP, are shown as circles. Interfamily comparisons are drawn with crosses.

variable in the different families. A certain number of connecting loops are exposed to the solvent, as is the helical region for its majority, while only the β-sheet is in contact with the multi domain assembly.

To investigate the structure and sequence similarity in the d.16 fold, the domains were structurally aligned with DALILite. In Fig. 2, a graph of the sequence alignment score, calculated as the raw sum of the BLOSUM62 (Henikoff and Henikoff, 1992) scores for the structurally equivalent positions, versus the DALI *Z*-score is reported for each pairwise comparison. As expected from the choice of ASTRAL/SCOP40 the sequence information is unable to clearly discriminate between couples of proteins belonging to the same or different families (see the zone of the graph with intermediate sequence scores). The structural score appears to be more effective in this task. However there are a set of intrafamily comparisons that have lower DALI *Z*-score than what is expected (in the range 3–7), and a small set of borderline interfamily comparisons with *Z*-scores approaching the cut-off of 9 that was proposed for identifying proteins belonging to the same family (Dokholyan *et al.*, 2002). This picture suggests a set of domains with a clear structural fingerprint, but also some cases that result difficult to be classified only on the basis of their structure.

Conformational sampling and essential dynamics

The conformational space of each of the 215 domains in the test set was sampled by CONCOORD (de Groot *et al.*, 1997). To assess the reliability of the obtained sampling, the CONCOORD ensembles were compared with the results from MD simulations. To this end a representative was chosen that better summarizes the secondary and tertiary features of each family in the d.16 fold. The selected candidates are highlighted in italics in Table II. Each domain was simulated for 16 ns (see Materials and methods section for details).

The RMSD and total energy statistics for the six representatives are reported in Table III. From both RMSD and Total Energy values it is evident that the proteins were stable during the simulations. RMSD graphs (data not reported) and visual inspection of the trajectory frames confirmed a general stability of the simulations, with an equilibration stage around 1 ns and a productive phase that encompassed

Table III. Summary results for MD simulations and CONCOORD runs

SCOP ID	PDB ID	MD					CONCOORD	MD	CONCOORD versus MD	CONCOORD versus B factor	MD versus B factor
		RMSD (nm)		Total energy (KJ/mol)		No. of eigenvectors	Explained variance (%)		RMSF correlation	RMSF correlation	RMSF correlation
		mean	SD	Mean	SD		Eigenvect. 1–3	Eigenvect. 1–8			
d.16.1.1	*1GPE*	*0.28*	*0.03*	*−393047*	*379*	*588*	*76.9*	*82.8*	*0.84*	*0.08*	*−0.23*
	1JU2					510	77.6			0.25	
	1KDG					543	69.7			0.28	
	1N4W					396	67.5			0.28	
d.16.1.2	*1K0I*	*0.39*	*0.07*	*−289570*	*328*	*306*	*78.8*	*85.2*	*0.83*	*−0.01*	*0.08*
	1PN0					273	74.5			0.57	
d.16.1.3	*1C0P*	*0.25*	*0.03*	*−215725*	*281*	*285*	*75.2*	*71.8*	*0.78*	*0.09*	*−0.19*
	1EL5					312	78.7			0.21	
	1NG4					264	72.2			0.56	
	1VE9					279	74.2			0.36	
d.16.1.5	*1S3E*	*0.29*	*0.05*	*−221441*	*286*	*336*	*76.3*	*81.0*	*0.85*	*0.16*	*0.05*
	1PJ5					357	78.8			0.47	
	1F8R					339	74.8			0.10	
	1SEZ					336	71.8			0.22	
	1B37					336	79.7			0.02	
d.16.1.6	1D5T	0.28	0.04	−199392	273	291	80.3	73.6	0.86	0.39	0.27
	1LTX					321	77.5			0.52	
d.16.1.7	1I8T	0.34	0.06	−216258	281	207	84.0	85.1	0.67	0.52	0.42

Representative domains for method assessment are in italics.

the remaining 15 ns. The thermal bath was effective in keeping the temperature around 300 K (data not reported).

Both the CONCOORD data and the MD trajectories were subjected to ED analysis to extract the information on the flexibility.

In the choice of the dimensionality of the essential subspace the number of directions to retain was chosen independently for the two methods and, in each method, the same number of directions was used for all the domains. As shown in Table III, the ensembles of 2000 structures generated by CONCOORD show a neater separation of the essential subspace, with displacements in the range 75–84%, already in the first three directions of motion, while the MD simulations require the inclusion of eight directions to explain a similar amount of conformational flexibility (displacements in the range 72–85%). This is a general result for CONCOORD runs: an ensemble of 500–1000 structures is usually enough to obtain a good sampling of the neighborhood of the starting structure and this is usually reflected by high values for the index of convergence (de Groot *et al.*, 1997).

The distribution of motion along the eigenvectors (data not shown) confirmed that all the excluded directions registered only small fluctuations associated with uninformative high frequency modes. The reduction was indeed effective in capturing the largest and most informative motions.

Flexibility analysis and representation

A residue-based description of the local flexibility was obtained by calculating the RMSF values for the positions of the Cα atoms. The analysis was performed on the structures after their projection into the essential subspace. Figure 3 reports the RMSF graphs for each family representative in the d.16 fold with the RMSF values from both the CONCOORD and MD ensembles. With the exception of the 1K0I domain, in all the cases along the MD trajectories there is a lower extent of fluctuation. Despite this, for all proteins, there is a good agreement in the location of the highest mobile peaks and the relative extents of residue flexibility.

The correlation coefficients between the RMSF vectors obtained by the two methods are reported in Table III. With the exception of 1I8T, the average correlation is 0.83, with a range of values that is higher than the one reported in the original CONCOORD paper (de Groot *et al.*, 1997). This better fitting can be the result of a considerably longer simulation time with respect to the original work and it gives an additional support to the reliability of CONCOORD as a conformational sampling method.

To compare flexibilities obtained from molecular simulations to those from crystallographic data, also the RMSF derived from the X-ray B-factors (see Materials and methods section) for the representative domains were included in Fig. 3 and the correlation coefficients with respect to RMSF values obtained by both the computational methods were collected in Table III. From the visual analysis, it appears that the B-factors describe an extremely reduced extent of fluctuation with respect to both the MD and the CONCOORD simulations for all the domains. Moreover, noticeable differences are observed also in the location and the relative height of some fluctuation peaks, particularly in the most flexible regions.

This lack of agreement is quantitatively confirmed by the correlation coefficients. These show a slightly better fit with the MD data, in some cases, and with the CONCOORD data, in others. On the whole, they indicate that poor (or none) correlation exists for the test set domains between RMSF from crystallography and those from simulation in solution. From this analysis, it is also expected that RMSF derived by the B-factors might not be able to discriminate flexibilities of different domains.

A further observation emerges by comparing the RMSF graphs from all the three methods with the secondary structure assignment (Kabsch and Sander, 1983) schematically reported in Fig. 3. As expected, some of the high fluctuation peaks obtained by simulation methods correspond to the loop regions. However, in many cases, also α helices as well as supersecondary structures including helices and loops are characterized by significant fluctuations. The mean flexibility of each secondary structure element (Table IV) indicates that, while B-factors describe a constrained and uniform distribution of residue fluctuations in the different secondary structure elements, simulation methods (and particularly CONCOORD) find that helices have a significant flexibility that is intermediate between the more mobile loops and the constrained strands. The expected relative order of flexibility associated to the secondary structure annotation is therefore reproduced by the mean and the median values of the calculated RMSF.

It has to be noted that the order of flexibility found for different secondary structure elements has not been introduced artificially by CONCOORD constraints. Indeed, the evaluation of correlation between flexibility and the number of constraints imposed on each residue (Supplementary Material, Table SI) shows lack of correlation between the two indexes. In this table, the correlation coefficients are reported for constraints that characterize the secondary structure class, according to CONCOORD 1.2 (i.e. tight phi/psi, loose phi/psi, secondary structure, sheet restrictions). Moreover, the program parameters that define the upper and lower distance limits for atoms in the same secondary structure element (obtained from MD analysis by de Groot and coworkers) have same magnitude for pairs of atoms in strands or helices.

Structural alignment of d.16 domains

Before performing the comparisons, the d.16 domains were structurally aligned with DALI (Holm and Sander, 1993), as it is acknowledged as one of the best performing structural alignment methods (Sierk and Pearson, 2004).

The domains in the test set show different degrees of structural equivalence when all-against-all pairwise comparisons are performed. For the majority of these, the two proteins have a large portion of the structure (higher than 60%) that is alignable. On the contrary, according to the SCOP classification, among all possible pairwise comparisons for proteins of the d.16 fold, there are more interfamily (84.3%) than intrafamily (15.7%) comparisons. Therefore, a weak correlation between the fraction of structurally aligned residues and the separation in families is observed. This is confirmed by Fig. 4 where, for each comparison, the percentages of alignment with respect to the two domains are reported. As expected, these values are high for both interfamily and intrafamily comparisons.

It has to be noted that for the intrafamily comparisons, beside an amount of alignable residues over 60%, it is

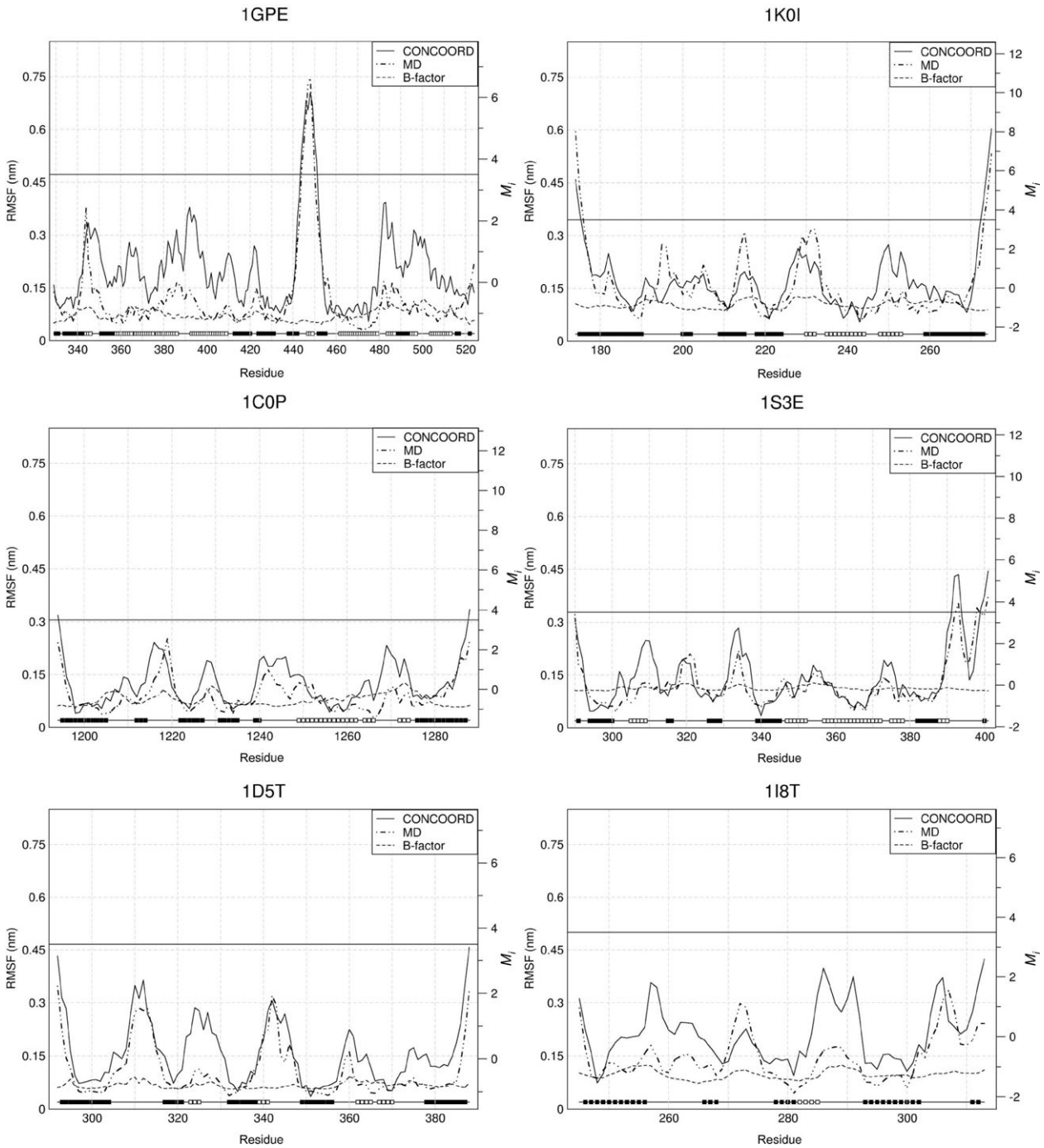

Fig. 3. RMSF profiles for the representative domain of each family in the d.16 fold as derived from MD simulations and CONCOORD runs (after reduction by ED analysis) and from B-factors. M_i values for the RMSF are reported on the secondary y-axis (solid line: $M_i = 3.5$). Residue numbering is reported according to the original PDB deposition. The secondary structure attribution for each residue is specified: black square = strand, white square = helix, solid line = loop.

evident that the two percentages tend to have similar values. This indicates that the superposition includes a similar fraction of the overall fold for each domain. On the contrary, in the interfamily pairs, there are both largely aligned structures and pairs of domains where one superimposes to only a part of the other, suggesting that across the superfamily there are structures that can be regarded as the extension of others from a common structural core. This is in agreement with a modular picture of the protein structure evolution (Kihara and Skolnick, 2003).

Flexibility comparison

For each pairwise comparison of RMSF vectors, a FC index was computed. After filtering on the basis of the structurally equivalent positions, the representations of the domain flexibilities are reduced to a pair of vectors of the same length. To measure the extent of similarity, it is therefore straightforward to compute their Pearson correlation coefficient. It is known that this index tends to overestimate correlation due to the presence of outliers that, in the case of flexibility profiles, are segments with unusually high flexibility (such

Table IV. Mean flexibility of secondary structure elements

Secondary structure	RMSF (nm)								
	CONCOORD			MD			B-factor		
	Mean	SD	Median	Mean	SD	Median	Mean	SD	Median
Strand	0.125	0.0725	0.104	0.104	0.0627	0.0863	0.0833	0.0247	0.0769
Helix	0.157	0.0900	0.135	0.104	0.0937	0.0850	0.0890	0.0259	0.0815
Loop	0.202	0.108	0.174	0.155	0.0966	0.132	0.0891	0.0273	0.0842

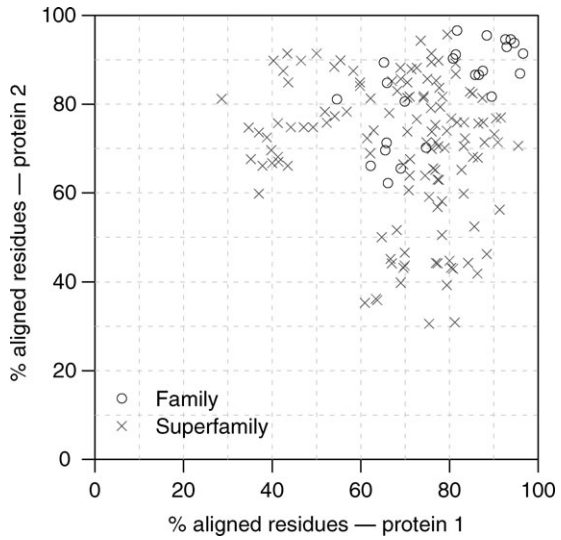

Fig. 4. Graph of the percentages of aligned residues for each pairwise comparison in the d.16 fold, as obtained by the DALI method. Comparisons between pairs of proteins belonging to the same family, as reported in SCOP, are shown as circles. Interfamily comparisons are drawn with crosses.

as protein ends and some loops) (Hunenberger *et al.*, 1995). Therefore, a robust median-based statistic (see Materials and methods section), proposed by Smith *et al.* (Smith *et al.*, 2003) to detect outliers in the B-factor distributions and successfully employed by other authors (Radivojac *et al.*, 2004), was applied to the RMSF vectors to identify and discard outliers before calculating the correlation coefficient. As evidenced by the line in Fig. 3, the M_i cutoff value of 3.5, proposed by Smith *et al.* (Smith *et al.*, 2003) on the basis of an extended statistics on flexibilities derived from B-factors, allows to eliminate the few very high peaks that could affect correlation (for example that around residues 440–460 in the 1GPE domain and the N- and C-terminal regions in the 1K0I). Moreover, it was verified that, in our test set, the choice of any M_i value in the range 3.0–4.0 does not affect significantly the obtained correlation coefficients.

The resulting values of flexibility correlation (FC index) are shown in Table V (in the lower half of the matrix), where the intrafamily comparisons are highlighted in italics. While the values range from 0.01 to 1.00, around 65% of the correlations are higher than 0.6, where only the 13.7% exceed 0.80. This suggests that there are extended similarities across the entire set, but the fraction of highly correlated domains is around the percentage of expected intrafamily comparisons (15.7%).

Looking at the correlations within the families, a general trend for stronger correlations than in the interfamily comparisons emerges. This is evident for the families 1 and 3, partially for the families 5 and 6, but not at all for family 2.

Table V. FC (lower half) and normalized RMSD (upper half) matrix from the DALILite alignment

		1				2		3				5					6		7
		1GPE	1JU2	1KDG	1N4W	1K0I	1PN0	1C0P	1EL5	1NG4	1VE9	1S3E	1PJ5	1F8R	1SEZ	1B37	1D5T	1LTX	1I8T
1	1GPE	—	*0.55*	*0.57*	*0.85*	0.81	0.85	0.93	0.74	0.73	0.76	0.82	0.74	0.89	0.76	0.72	0.69	0.68	0.65
	1JU2	*0.47*	—	*0.71*	*0.84*	0.72	0.78	0.84	0.88	0.78	0.71	0.81	0.86	0.83	1.00	0.81	0.82	0.80	0.63
	1KDG	*0.47*	*0.74*	—	*0.87*	0.83	0.97	0.86	0.69	0.73	0.83	0.81	0.84	0.82	0.91	0.83	0.68	0.77	0.80
	1N4W	*0.72*	*0.75*	*0.68*	—	0.74	0.76	0.79	0.83	0.75	0.76	0.80	0.80	0.84	0.76	0.87	0.68	0.80	0.84
2	1K0I	0.37	0.70	0.19	0.31	—	*0.55*	0.69	0.78	0.80	0.81	0.78	0.66	0.79	0.76	0.80	0.69	0.84	0.50
	1PN0	0.36	0.64	0.56	0.68	*0.55*	—	0.65	0.72	0.75	0.74	0.72	0.72	0.65	0.66	0.77	0.74	0.72	0.79
3	1C0P	0.62	0.67	0.53	0.63	0.47	0.85	—	*0.63*	*0.42*	*0.34*	0.67	0.53	0.82	0.64	0.76	0.63	0.70	0.58
	1EL5	0.08	0.54	0.63	0.48	0.41	0.81	*0.90*	—	*0.45*	*0.58*	0.74	0.49	0.73	0.63	0.72	0.70	0.69	0.56
	1NG4	0.39	0.72	0.65	0.48	0.47	0.64	*0.84*	*0.88*	—	*0.34*	0.78	0.54	0.84	0.70	0.76	0.55	0.60	0.46
	1VE9	0.32	0.61	0.66	0.67	0.58	0.82	*0.87*	*0.85*	*0.82*	—	0.80	0.52	0.71	0.83	0.78	0.63	0.66	0.63
5	1S3E	0.49	0.63	0.69	0.59	0.50	0.74	0.85	0.78	0.62	0.81	—	*0.70*	*0.43*	*0.59*	*0.50*	0.61	0.59	0.59
	1PJ5	0.40	0.68	0.51	0.29	0.36	0.64	0.60	0.77	0.64	0.69	*0.51*	—	*0.67*	*0.64*	*0.66*	0.68	0.65	0.97
	1F8R	0.40	0.71	0.42	0.65	0.57	0.71	0.75	0.52	0.73	0.71	*0.90*	*0.45*	—	*0.56*	*0.49*	0.59	0.58	0.66
	1SEZ	0.31	0.61	0.56	0.61	0.22	0.73	0.71	0.69	0.73	0.70	*0.77*	*0.69*	*0.69*	—	*0.56*	0.63	0.78	0.50
	1B37	0.31	0.39	0.42	0.53	0.26	0.64	0.57	0.75	0.62	0.63	*0.81*	*0.56*	*0.79*	*0.71*	—	0.66	0.54	0.66
6	1D5T	0.39	0.73	0.67	0.74	0.26	0.75	0.81	0.81	0.74	0.75	0.88	0.75	0.77	0.77	0.66	—	*0.51*	0.68
	1LTX	0.19	0.83	0.63	0.65	0.50	0.83	0.77	0.68	0.59	0.72	0.80	0.74	0.80	0.89	0.65	0.87	—	0.65
7	1I8T	0.40	0.62	0.69	0.62	0.01	0.58	0.60	0.70	0.63	0.67	0.67	0.33	0.59	0.43	0.68	0.76	0.58	—

Intrafamily comparisons are in italics.

While the latter seems to be a case of effectively uncorrelated dynamics, that represents a single exception within this superfamily, in family 5 it looks as if the domains 1PJ5 and 1SEZ show a dynamics different from those of the other three strictly related domains. There are also some sparse cases of highly correlated interfamily pairs.

The overall picture is in good agreement with the SCOP classification, and suggests the presence of a distinguishable familiar trend in the flexibility profiles, that can be viewed as the 'dynamical fingerprint' of the family.

A deeper insight into the ability of the FC index to detect intrafamily relationships can be obtained by comparing the FC index with an index of structural similarity. For this analysis, the simple RMSD estimate was preferred to the DALI *Z*-score on the basis of the observation that the latter represents a statistical index based on a random background distribution and consequently it cannot be directly compared to the FC index. The normalized RMSD values are reported in the upper half of the matrix in Table V. It clearly emerges that the FC and the RMSD indexes have different informativeness. While the FC index shows the best discrimination ability in identifying the pairs of domains belonging to families 1 and 3, the RMSD clearly identifies lower intra than interfamily distances for the families 2 and 6, partially discriminates the 3 and 5 families and fails in family 1, whose domains are structurally more similar to domains belonging to other families.

An exploratory cluster analysis (see Materials and methods section) was also applied to the complement of the FC matrix and, for comparative purposes, to the normalized RMSD matrix. The resulting dendrograms are reported in Fig. 5a and b, respectively. As expected on the basis of the results in Table V, the FC index is able to cluster family 1, 3 and 6, while it shows some difficulties in grouping the 1PJ5 and 1SEZ domains with the others of family 5 and fails in clustering family 2. On the other hand, the RMSD well identifies families 2 and 6, partially families 3 and 5 and fails with family 1.

To have a rough insight into the effects of combining the two indexes, the cluster analysis was applied to the linear combination of the normalized RMSD and the complement to FC index (coefficients 0.5). This combination gives the best results, with only a single misplacement for the protein 1PJ5 from family 5 (Fig. 5c).

Despite semiquantitative, this other result suggests that employment of the dynamical information is complementary to structural comparison in detecting similarities among distant homologous proteins.

A possible drawback in our analysis could be the dependence of the results from the choice of a particular method for the structural alignment. To verify this point, the same procedure was applied to the d.16 set to derive the FC indexes, by using the Structal method (Gerstein and Levitt, 1998) to align the domains before pairwise comparisons. At difference with DALI that directly search for a good alignment, this method searches for transformations that optimally position the two structures with respect to one another and then use the transformation to find the best alignment. The resulting FC data and, for comparison, the normalized RMSD values are shown in Table VI. These values and

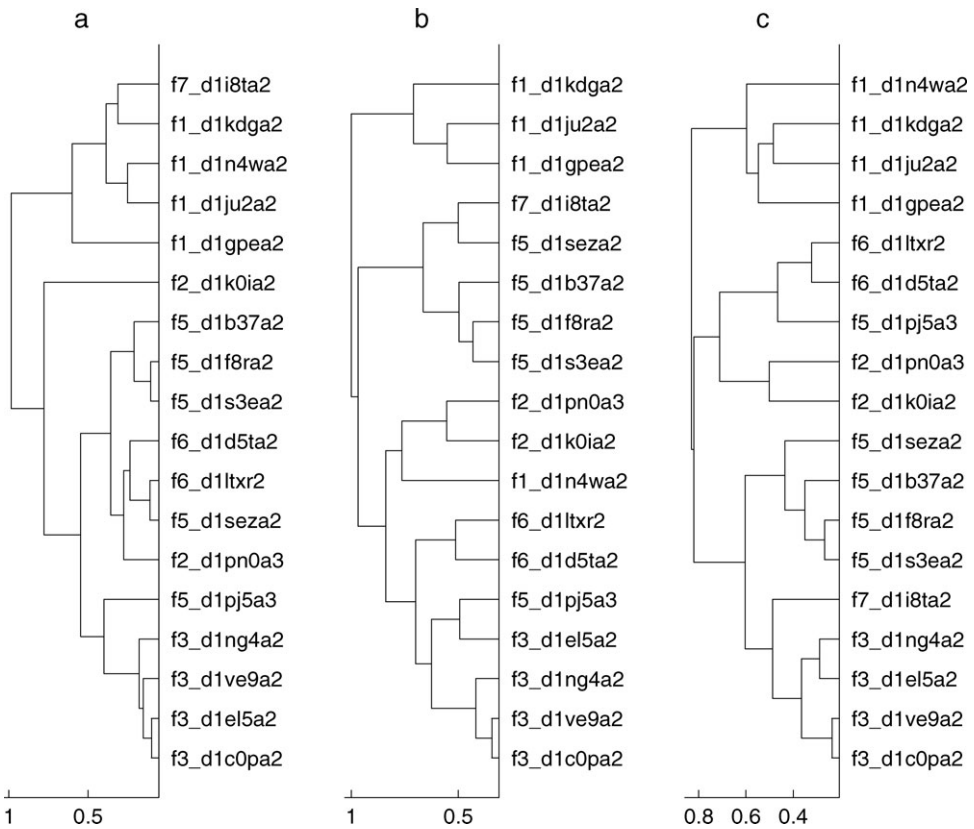

Fig. 5. Dendrograms from cluster analysis on the d.16 fold. Structural alignment obtained by DALI. (**a**) Results from the complement to the FC matrix; (**b**) results from the normalized RMSD; (**c**) results from a linear combination of the two indexes with equal coefficients (0.5).

Table VI. FC (lower half) and normalized RMSD (upper half) from the Structal alignment

		1				2		3				5					6		7
		1GPE	1JU2	1KDG	1N4W	1K0I	1PN0	1C0P	1EL5	1NG4	1VE9	1S3E	1PJ5	1F8R	1SEZ	1B37	1D5T	1LTX	1I8T
1	1GPE	—	*0.25*	*0.30*	*0.66*	0.53	0.46	0.63	0.52	0.55	0.60	0.96	0.56	1.07	0.52	0.90	0.51	0.58	0.41
	1JU2	*0.46*	—	*0.41*	*0.51*	0.97	0.67	0.47	0.75	0.51	0.49	0.63	0.54	0.52	1.00	0.74	0.48	0.59	0.47
	1KDG	*0.44*	*0.77*	—	*0.63*	0.55	0.68	0.54	0.50	0.52	0.58	0.65	0.51	0.43	0.62	0.87	0.62	0.51	0.54
	1N4W	*0.68*	*0.73*	*0.69*	—	1.00	0.67	0.82	0.64	0.56	0.61	0.73	0.65	0.43	0.45	0.52	0.79	0.54	0.53
2	1K0I	0.36	0.45	0.44	0.43	—	*0.27*	0.66	0.76	0.46	0.74	0.69	0.81	0.48	0.40	0.71	0.56	0.69	0.30
	1PN0	0.46	0.51	0.51	0.56	*0.60*	—	0.47	0.43	0.51	0.59	0.74	0.48	0.74	0.41	0.85	0.74	0.42	0.46
3	1C0P	0.07	0.48	0.33	0.53	0.54	0.66	—	*0.58*	*0.18*	*0.19*	0.55	0.36	0.57	0.45	0.74	0.51	0.59	0.61
	1EL5	0.44	0.34	0.64	0.37	0.38	0.83	*0.86*	—	*0.45*	*0.41*	0.61	0.35	0.76	0.51	0.79	0.77	0.42	0.62
	1NG4	0.35	0.63	0.33	0.39	0.50	0.39	*0.85*	*0.87*	—	*0.33*	0.50	0.38	0.54	0.49	0.50	0.33	0.45	0.36
	1VE9	0.33	0.62	0.51	0.18	0.55	0.58	*0.89*	*0.84*	*0.82*	—	0.55	0.34	0.61	0.52	0.75	0.51	0.48	0.63
5	1S3E	0.42	0.41	0.48	0.58	0.49	0.62	0.87	0.78	0.57	0.83	—	*0.63*	*0.44*	*0.60*	*0.36*	0.28	0.55	0.64
	1PJ5	0.31	0.72	0.58	0.31	0.41	0.31	0.56	0.75	0.73	0.64	*0.70*	—	*0.50*	*0.45*	*0.58*	0.49	0.51	0.66
	1F8R	0.14	0.34	0.48	0.76	0.63	0.61	0.75	0.59	0.65	0.67	*0.88*	*0.49*	—	*0.33*	*0.34*	0.37	0.63	0.34
	1SEZ	0.08	0.28	0.40	0.56	0.41	0.70	0.85	0.73	0.69	0.79	*0.72*	*0.45*	*0.69*	—	*0.61*	0.74	0.51	0.33
	1B37	0.32	0.49	−0.03	0.45	0.38	0.60	0.44	0.65	0.52	0.54	*0.82*	*0.57*	*0.80*	*0.57*	—	0.42	0.60	0.32
6	1D5T	0.16	0.55	0.65	0.72	0.46	0.71	0.85	0.67	0.74	0.85	0.47	0.74	0.76	0.74	0.73	—	*0.28*	0.45
	1LTX	0.17	0.64	0.63	0.74	0.23	0.57	0.72	0.53	0.44	0.63	0.78	0.50	0.80	0.77	0.67	*0.88*	—	0.50
7	1I8T	0.39	0.65	0.68	0.26	0.32	0.59	0.54	0.52	0.45	0.53	0.69	0.43	0.49	0.49	0.64	0.39	0.33	—

Intrafamily comparisons are in italics.

the corresponding dendrograms from cluster analysis (Fig. 6a and b) evidence that the FC index maintains its ability to cluster families 1, 3 and 6, and also the difficulties in grouping the 1PJ5 and 1SEZ domains with the others of family 5, as observed in the data obtained by using the DALI alignment. On the other hands, from this analysis it emerges that the RMSD obtained by the Structal alignments have a reduced ability to cluster domains belonging to the same family. Therefore the different informativeness of the two indexes (FC and RMSD) emerges more clearly than in the analysis based on the DALI alignment. As a consequence, also the linear combination of the two indexes shows a discrimination ability similar to that of the FC index alone, as shown by the dendrogram in Fig. 6c.

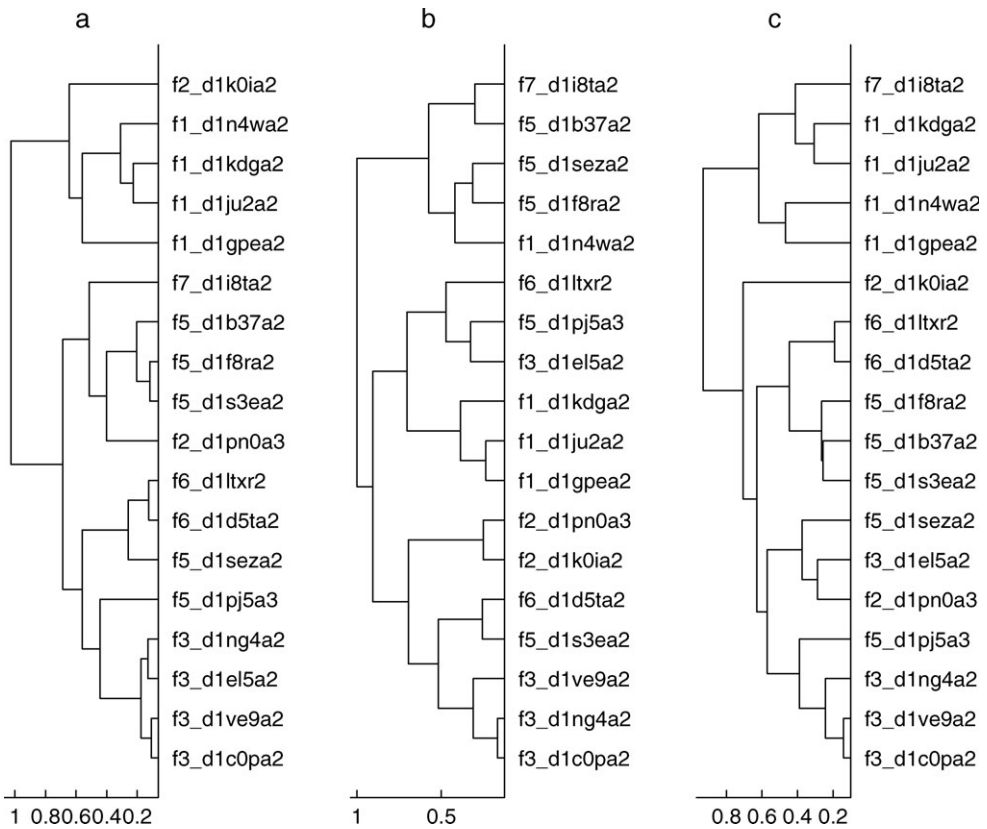

Fig. 6. Dendrograms from cluster analysis on the d.16 fold. Structural alignment obtained by Structal. (**a**) Results from the complement to the FC matrix; (**b**) results from the normalized RMSD; (**c**) results from a linear combination of the two indexes with equal coefficients (0.5).

Table VII. Analysis of performance: AUC values from ROC curve analysis

Alignment	Assessment	Index	SCOP ID							
			d.3	d.14	d.16	d.79	d.142	d.144	d.153	d.169
DALI	Family	RMSD	0.91	0.91	0.91	0.97	0.79	0.94	0.95	0.97
		FC	0.79	0.74	0.87	0.84	0.94	0.96	0.94	0.83
	Superfamily	RMSD	0.68	0.85	0.81	0.94	0.71	0.77	0.89	0.91
		FC	0.68	0.61	0.76	0.83	0.85	0.90	0.92	0.79
Structal	Family	RMSD	0.96	0.94	0.96	0.99	0.92	0.96	0.99	0.98
		FC	0.79	0.69	0.89	0.73	0.93	0.95	0.94	0.83
	Superfamily	RMSD	0.70	0.92	0.90	0.98	0.86	0.88	0.98	0.97
		FC	0.64	0.57	0.73	0.81	0.79	0.90	0.91	0.78

Significance

The statistical significance of the d.16 analysis was assessed by employment of the *K* correlation index (see Materials and methods section for details). This index returns a quantitative, synthetic and accurate measure of the extent of correlation in the matrix. The value obtained for the FC matrix based on the DALI structural alignment (Table V) is 60.7% and for that based on the Structal alignment (Table VI) is 53.0%. This confirms a trend of extended similarities in the dynamic features of the set. This was not obtained by chance, as demonstrated by a random test, were a distribution of 1500 random comparisons were generated by shuffling the values in the original RMSF vectors, after the structural alignment and vector filtering. With this approach, the set of 1500 similarity matrices was representative of the random background distribution for flexibility comparisons among a generic collection of proteins with the same structural alignments of the 18 in the test case. The *K* index was calculated on each matrix and the *K* average value was $21.2 \pm 1.3\%$. A consequence of this result is that the FC index can be viewed as a quite informative measure of similarity.

In a second test, the *K* values for the FC matrices were compared with an artificial model of optimal separation, built on the basis of the SCOP classification. The model similarity matrix was constructed by placing 1.0 on the diagonal, 0.85 for each comparison of two proteins from the same family (given the observed distribution of inter and intrafamily relationships within the d.16 SCOP fold) and 0.0 for all the other comparisons. The *K* value obtained for this model is 60.0%. This indicates that the extent of similarity described by the FC index is in quite good agreement with what is needed to obtain the separation in families as reported in SCOP.

Analysis of performance

To assess the performance of the proposed procedure for flexibility comparison, a large scale analysis was performed on the entire test set (Table I) by calculating the receiver operating characteristic (ROC) curves (see Materials and methods section) for the FC index. The performance was also compared to that of the RMSD on equivalent Cα atoms, as identified by DALI and Structal. This structure comparison index was chosen instead of the statistical indexes based on random background distributions (the DALI *Z*-score or the Structal *P*-value) to allow a direct comparability to the FC index which is not a standardized measure. The discrimination abilities of the two indexes were evaluated independently for the two SCOP classification levels of family and superfamily, and the results are summarized in Table VII by means of the synthetic area under the curve (AUC) index.

It can be observed that the AUC values for the FC index, AUC(FC), for each fold, derived either from the DALI or the Structal alignment, are very similar. On the contrary, the AUC(RMSD) values show a significant dependence on the aligment method. In particular, accuracy differences up to the 15% are observed for the attribution of domains to superfamilies. This suggests the independence of the results of the flexibility comparison from the structural alignment method.

In the comparison of the FC to the RMSD, it can be observed that the relative performance of the two indexes varies among the different folds. This results clearer by examining the ROC curves for the two indexes, based for example on the DALI structural alignment, as shown in Fig. 7. In the figure, the curves for the eight SCOP folds are grouped in three graphs, according to the relative degree of accuracy of the indexes, as evaluated by the AUC values reported in Table VII. Both in the classification of domains belonging to different families (top graphs in Fig. 7) and to different superfamilies (bottom graphs) the FC index outperforms the RMSD for some folds, shows a performance similar to that of the RMSD for others, and results less performant for the remaining. In general, the FC index exploits a greater accuracy in discriminating domains belonging to different superfamilies than to different families. The flexibility information appears to have a major role in discriminating the domains belonging to the d.142 and the d.144 folds (the ATP-grasp and the Protein kinase-like folds, respectively). For the d.16 fold, the comparison of the ROC curves confirmed that the two indexes have a similar discriminating ability (with AUC values that differ $<10\%$).

Discussion

In this work, the informativeness of intrinsic protein flexibility in the detection of similarities among distant homologous proteins was investigated.

In consequence of limitations arising from the employment of experimental flexibilities, molecular simulations were chosen as the source of dynamical information and, in view of large-scale applications, fast conformational sampling methods were preferred.

Among those, a quite promising one is CONCOORD, a fast and efficient method to generate ensembles of

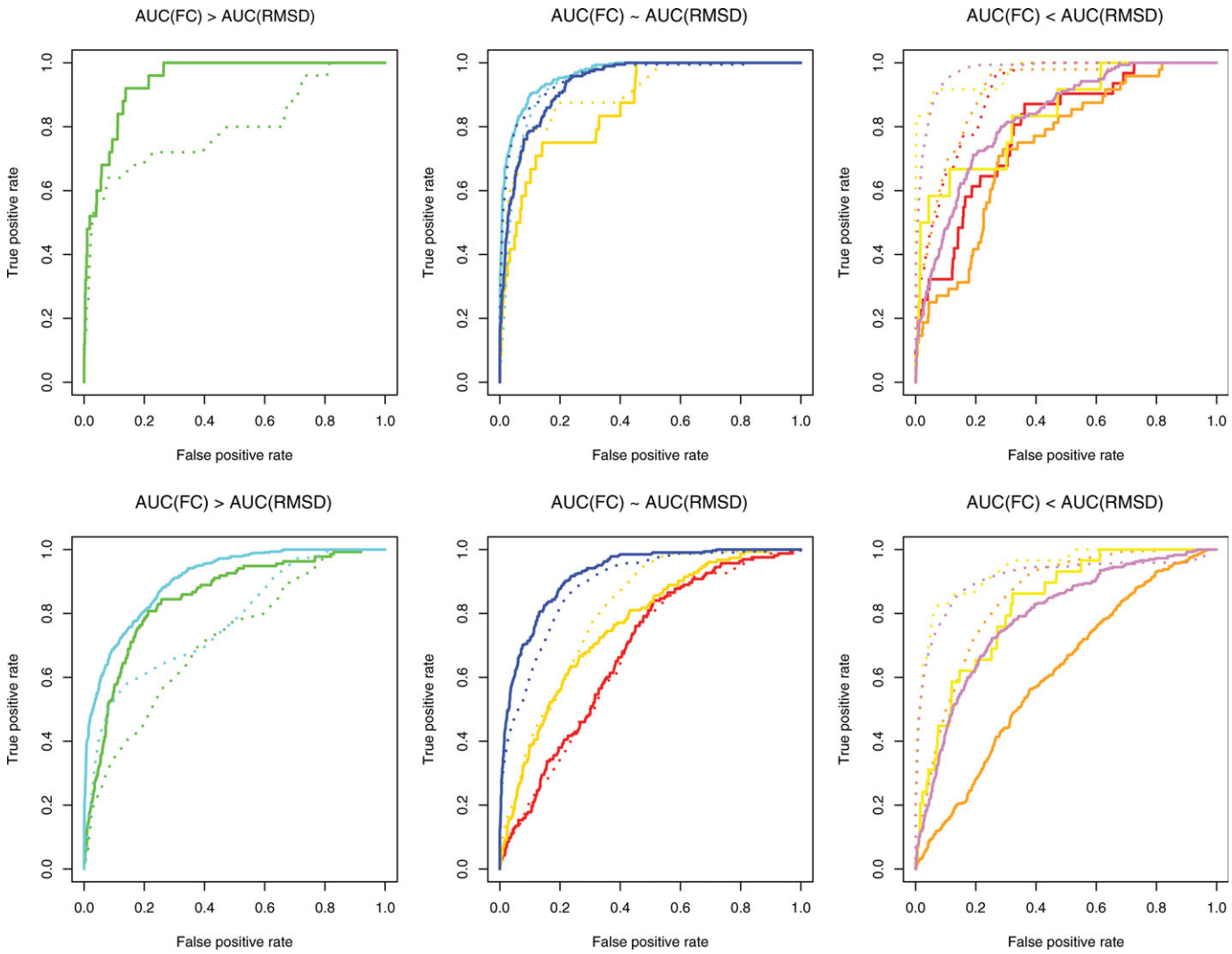

Fig 7. ROC curves for the search results with the FC index (solid lines) and the RMSD from the DALI structural alignments (dotted lines) for the 8 folds in the test set: d.3, red; d.14, orange; d.16, gold; d.79, yellow; d.142, green; d.144, cyan; d.153, blue; d.169, purple. Top: intrafamily comparisons; bottom: interfamily comparisons. From left to right: ROC curves with AUC(FC) greater that, equal to and less than AUC(RMSD) values (see Table VII).

independent structures (Tai, 2004; de Groot *et al.*, 1997). Application studies reported the assessment of the method and demonstrated its reliability for single cases of different types of proteins (de Groot *et al.*, 1997; Kleinjung *et al.*, 2000; Barrett *et al.*, 2004). In this work, the reliability of the method was further investigated: on one side, for the first time, the assessment was extended to a superfamily with a significant number of proteins, on the other side, this was done by comparison with long MD simulations. The results highlighted a good agreement between the RMSF profiles derived from MD and CONCOORD, demonstrating that the per-residue flexibilities are consistent in location and relative amplitude (Fig. 3 and Table III). This suggests that CONCOORD can effectively capture the essential feature of the protein flexibilities with accuracy comparable to traditional long MD simulations but with reduced computational effort; this fact also supports its employment for a large-scale annotation of the protein flexibility.

A further confirmation of the choice of employing flexibilities from molecular simulations instead of crystallographic data arose from the comparison of the RMSF profiles derived from the simulations and B-factors. The noticeable differences that emerged in the RMSF profiles as well as the poor correlation coefficients (Fig. 3 and Table III) suggested a minor ability of the B-factors in discriminating the flexibilities of the different domains. In fact, the d.16 fold dendrogram obtained by cluster analysis on the FC matrix derived from the B-factors confirmed the lack of discrimination ability of this source (Supplementary Material, Figure S1).

On the other hand, the secondary structure annotation is commonly used for describing the local flexibility of a protein structure, as the different constraints imposed by the H-bonds network to the different secondary structure elements can indeed be associated to different degrees of residue fluctuation (Andersen *et al.*, 2002). We proved that the mean fluctuations obtained by molecular simulations, and in particular by CONCOORD, do reproduce the relative order that is expected on the basis of secondary structure annotation (Table IV). However, the single RMSF profiles (Fig. 3) demonstrated that a more detailed description of the relative flexibilities of the different domain regions is introduced by employment of simulated flexibilities.

As the calculated RMSF values can be regarded as a vector of flexibility in the space of the protein residues,

protein–protein comparison can be performed by measuring the similarity between flexibility vectors instead of the one between amino acid sequences or tertiary structures. The vectorial representation looks suitable for the classical vector operations such as the normalized internal product (the Pearson correlation coefficient). This is a widely employed measure of similarity that, with an appropriate filter for outliers (Smith *et al.*, 2003), can be easily applied to vectors with values on different scales and therefore allows to compare RMSF for proteins with different global amplitude of motion.

Because equivalent flexible regions are often located in structurally equivalent parts of the proteins, the structural alignment can be the appropriate framework to compare protein dynamics (Pandini and Bonati, 2005). On this basis, the structural alignment was employed to reduce the vectors by selecting only the corresponding residues. Consequently, the correlation coefficient of a pair of equivalent vectors gave a synthetic measure of the extent of similarity between the average motions of the two proteins.

The reduction due to the alignment did not affect dramatically the original information because a large part of the vectors were depleted of $<30\%$ of their length. This guaranteed that the procedure was still performing a domain–domain comparison and it was not reduced to the comparison of the flexibility of local structures. Moreover, the relatively low depletion guaranteed that, after filtering, the result of the initial PCA was not distorted and the definition of the essential subspace preserved. It was also verified that the amount of explained variance along each principal component was evenly reduced by filtering and the order of the eigenvectors maintained (results not shown). Additionally, it should be noted that the non-equivalent positions in the structural alignments did not belong preferentially to loop regions, as it could be expected on the basis of their usual higher divergence across the evolution; this is a remarkable result because it is generally expected that the high loop flexibilities do contribute to a large extent to the most informative directions of motion, and it further confirms that the excluded residues do not affect significantly the definition of the essential subspace.

The inclusion of a structural alignment step may anyway constitute a possible drawback in the procedure. First, the results obtained may be influenced by the choice of a particular method for the structure alignment. Second, there is the risk to insert a source of information from the structural alignment in the model and therefore to add a bias to the similarity index, or in the worst case, to reduce the flexibility comparison to a structural comparison.

The first point was addressed by comparing the FC data obtained for the d.16 fold on the basis of the DALI (Table V and Fig. 5a) and the Structal (Table VI and Fig. 6a) alignments. The analysis showed a comparable ability of the FC index in grouping domains belonging to the same family, independently of the clustering ability of the corresponding structural alignment method, as described by the RMSD index (Figs 5b and 6b). More interestingly, the analysis of the performances of the FC index on the entire test set demonstrated that the accuracy of this index in discriminating domains both at the family and the superfamily levels, as synthetized by the AUC values, is conserved for the two structural alignment methods (Table VII).

The question of the independent informativeness of the flexibility index with respect to structural similarity indexes can be simply confuted by comparing the flexibility index and the RMSD for the d.16 set of pairwise estimates. The correlation coefficient between the two indexes was -0.31 for the analysis based on the DALI alignment (-0.33 for that based on the Structal alignment).

It was hypothesized that an additional piece of information from dynamics analysis could indeed arise from the functional similarities, often partially hidden by a static comparison. This was demonstrated, in the case of families belonging to the d.16 fold, by the discrimination ability of the FC index that resulted particularly effective for the comparisons involving families 1 and 3. In fact, a deeper analysis of the biological function of d.16 proteins shed light into the reasons of a different degree of discrimination across the families belonging to this fold. For domains belonging to the GMC oxidoreducatases (family 1) and the D-aminoacid oxidase-like (family 3) families a functional activity located in the d.16 domain has been reported. For the cholesterol oxidase, belonging to the family 1, two loops have been identified which act as a 'lid' over the active site facilitating binding of the substrates, and one of these loops is included in the d.16 domain. Interestingly, the differences observed in these loops' flexibilities among proteins of different species appear to translate into differences in substrate activity and specificity, despite a high structural conservation of the active site (Yue *et al.*, 1999). Also, for the D-aminoacid oxidases of the family 3 (1C0P and 1VE9), a loop located within the d.16 domain has been devised as a 'lid' controlling the active-site accessibility and plasticity as well as the increasing hydrophobicity of the cavity in the 'closed' conformation (Todone *et al.*, 1997; Pilone, 2000). Conversely, in most of the L-aminoacid/polyamine oxidase (family 5), the entrance cavity is located at the opposite end of a long and narrow channel leading to the catalytic site (Pawelek *et al.*, 2000; Binda *et al.*, 2001; Binda *et al.*, 2002; Edmondson *et al.*, 2004) and therefore the loop or helices' mobility associated to the substrate admission lies far from the d.16 domain. Similarly, for the FAD-containing aromatic hydroxylases (family 2), a complex catalytic mechanism has been devised (Enroth *et al.*, 1998; Ballou *et al.*, 2005) that involves large conformational changes in both FAD and protein regions external to the d.16 domain to allow substrates' access to the active site.

The resulting picture suggested a hypothesis on the informativeness of the domain flexibility. When the biologically relevant dynamics is embedded in the domain unit, the flexibility is more informative and the natural selective pressure leads to its conservation in the family. When this is not the case, the protein accepts some 'mutations' in its 'flexibility code'. An example can be protein 1PJ5 that exploits a dynamics different from those of the other domains in the family 5 (see the FC indexes in Table V and Fig. 5a) leading to a less satisfying detection of similarities within the family. This supports the hypothesis of a fine-tuned conservation of the dynamics that is inscribed in the sequence.

The comparative analysis of dendrograms from cluster analysis on the FC and RMSD indexes obtained for the d.16 fold (Figs 5 and 6) also supported the complementarity of the dynamical and structural information. In fact, for families where a partial discrimination ability is just exploited by

the RMSD, the FC index seemed to better group the domains into the family. Conversely, where a poor structural similarity as well as a poor clustering of the RMSD was observed among the domains, the FC index was able to improve the domains' clustering.

Thus, the connection between flexibility and functional similarities is evident in this test case, where distant homologous domains with also remarkable structural diversity are correctly assigned to families by the FC index. To this extent, flexibility annotation appears to be a promising tool to support an automatic functional annotation in those cases where a manual annotation is otherwise needed.

While on the d.16 fold the investigation was done for each step of the procedure, the general performance of the FC index was evaluated for a larger collection of proteins from different folds (Table I). This allowed to extend the test to assess the ability to discriminate similarities both at the family and the superfamily levels.

The ROC curves confirmed an intrinsic discrimination power of the FC index. This was slightly more efficient at the superfamily level, suggesting that the role of dynamics information can be of interest also at longer evolutionary distances. In agreement with the detailed result on d.16 fold, the FC index had, for some cases, a performance similar to that of the structural comparison index, whereas it was more effective than a geometrical comparison for other cases (the d.142 and the d.144 folds). It is expected that the degree of relative accuracy would increase with an increasing role of the flexibility in the biological activity.

In conclusion, comparison of domain flexibilities highlighted that dynamics may contribute to the detection of similarities of distant homologous proteins. The results suggested that flexibility can be regarded as complementary to structural information and that it plays its major role when the most informative motions detected within the domain have a specific functional role. This additional level of information, inaccessible by simple structural comparison, can be employed to detect functional similarities otherwise unrecoverable.

Due to the computational cost of MD simulations, the proposed procedure has been designed on a fast sampling method, but the FC index can be easily calculated from MD trajectory data as well. This suggests the application of this comparative approach to simulation databases when these will be available (Tai *et al.*, 2004).

Future directions opened by this work include, on one side, the extension of this study to collect a proper statistics and provide a standardized FC index and, on the other side, the direct employment of the information about residue flexibility to annotate sequences with known structure and then to search the sequence space with this additional feature.

An additional development would be the search of relationships between the regions with high degree of flexibility and the occurrences in that regions of some characteristic local structures to identify functional motifs.

Supplementary data

Supplementary data are available at *PEDS* online.

Acknowledgments

We are grateful to Dr Franca Fraternali and Dr Jens Kleinjung for critical reading of the manuscript and valuable discussions. We also thank Prof. Roberto Todeschini for useful suggestions regarding statistical indexes.

References

Amadei,A., Linssen,A.B. and Berendsen,H.J. (1993) *Proteins*, **17**, 412–425.
Andersen,C.A.F., Palmer,A.G., Brunak,S. and Rost,B. (2002) *Structure*, **10**, 175–184.
Bahar,I., Atilgan,A.R. and Erman,B. (1997) *Fold Des.*, **2**, 173–181.
Ballou,D.P., Entsch,B. and Cole,L.J. (2005) *Biochem. Biophys. Res. Commun.*, **338**, 590–598.
Barrett,C.P., Hall,B.A. and Noble,M.E.M. (2004) *Acta Crystallogr. D Biol. Crystallogr.*, **60**, 2280–2287.
Berendsen,H.J.C., Postma,J.P.M., van Gunsteren,W.F. and Hermans,J. (1981) In Pullman,B. (ed.), *Intermolecular Forces.* Dordrecht, Reidel, pp. 331–342.
Berendsen,H.J.C., van der Spoel,D. and van Drunen,R. (1995) *Comp. Phys. Comm.*, **91**, 43–56.
Binda,C., Angelini,R., Federico,R., Ascenzi,P. and Mattevi,A. (2001) *Biochemistry*, **40**, 2766–2776.
Binda,C., Mattevi,A. and Edmondson,D.E. (2002) *J. Biol. Chem.*, **277**, 23973–23976.
Brenner,S.E., Koehl,P. and Levitt,M. (2000) *Nucleic Acids Res.*, **28**, 254–256.
Chandonia,J., Hon,G., Walker,N.S., Lo Conte,L., Koehl,P., Levitt,M. and Brenner,S.E. (2004) *Nucleic Acids Res.*, **32**, D189–D192.
Chandonia,J., Walker,N.S., Lo Conte,L., Koehl,P., Levitt,M. and Brenner,S.E. (2002) *Nucleic Acids Res.*, **30**, 260–263.
Darden,T., York,D. and Pedersen,L. (1993) *J. Chem. Phys.*, **98**, 10089–10092.
de Groot,B.L., van Aalten,D.M.F., Scheek,R.M., Vriend,G. and Berendsen,H.J.C. (1997) *Proteins*, **29**, 240–251.
DeLano,W.L. (2002) *DeLano Scientific*. San Carlos, CA, USA.
Dokholian,N.V., Shakhnovich,B. and Shakhnovich,E.I. (2002) *Proc. Natl. Acad. Sci. USA*, **99**, 14132–14136.
Eastman,P., Pellegrini,M. and Doniach,S. (1999) *J. Chem. Phys.*, **110**, 10141–10152.
Edmondson,D.E., Mattevi,A., Binda,C., Li,M. and Hubalek,F. (2004) *Curr. Med. Chem.*, **11**, 1983–1993.
Enroth,C., Neujahr,H., Schneider,G. and Lindqvist,Y. (1998) *Structure*, **6**, 605–617.
Everitt,B. (1974) *Cluster analysis.* Heinemann Educational books, London.
Feenstra,K.A., Hess,B. and Berendsen,H.J.C. (1999) *J. Comp. Chem.*, **20**, 786–798.
Fraaije,M.W. and Mattevi,A. (2000) *Trends Biochem. Sci.*, **25**, 126–132.
Frauenfelder,H., Petsko,G.A. and Tsernoglou,D. (1979) *Nature*, **280**, 558–563.
Gerstein,M., Lesk,A.M. and Chothia,C. (1994) *Biochemistry*, **33**, 6739–6749.
Gerstein,M. and Levitt,M. (1998) *Prot. Sci.*, **7**, 445–456.
Gribskov,M. and Robinson,N.L. (1996) *Comput. Chem*, **20**, 25–33.
Grottesi,A. and Sansom,M.S.P. (2003) *FEBS Lett.*, **535**, 29–33.
Halle,B. (2002) *Proc. Natl. Acad. Sci. USA*, **99**, 1274–1279.
Henikoff,S. and Henikoff,J.G. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 10915–10919.
Hunenberger,P.H., Mark,A.E. and van Gunsteren,W.F. (1995) *J. Mol. Biol.*, **252**, 492–503.
Hess,B., Bekker,H., Berendsen,H.J.C. and Fraaije,J.G.E.M. (1997) *J. Comp. Chem.*, **18**, 1463–1472.
Holm,L. and Park,J. (2000) *Bioinformatics*, **16**, 566–567.
Holm,L. and Sander,C. (1993) *J. Mol. Biol.*, **233**, 123–138.
Hunenberger,P.H., Mark,A.E. and van Gunsteren,W.F. (1995) *J. Mol. Biol.*, **252**, 492–503.
Iglewicz,B. and Hoaglin,D.C. (1993) *How to detect and handle outliers.* ASQ Quality Press, Milwaukee, WI.
Kabsch,W. and Sander,C. (1983) *Biopolymers*, **22**, 2577–2637.
Kihara,D. and Skolnick,J. (2003) *J. Mol. Biol.*, **334**, 793–802.
Kleinjung,J., Bayley,P. and Fraternali,F. (2000) *FEBS Lett.*, **470**, 257–262.
Lindahl,E., Hess,B. and van der Spoel,D. (2001) *J. Mol. Mod.*, **7**, 306–317.
Maguid,S., Fernandez-Alberti,S., Ferrelli,L. and Echave,J. (2005) *Biophys. J.*, **89**, 3–13.
Meinhold,L. and Smith,J.C. (2005) *Biophys. J.*, **88**, 2554–2563.
Miura,R. (2001) *The Chemical Record*, **1**, 183.

Murzin,A., Brenner,S., Hubbard,T. and Chothia,C. (1995) *J. Mol. Biol.*, **247**, 536–540.
Pandini,A. and Bonati,L. (2005) *Protein Eng. Des. Sel.*, **18**, 127–137.
Pawelek,P.D., Cheah,J., Coulombe,R., Macheroux,P., Ghisla,S. and Vrielink,A. (2000) *EMBO J.*, **19**, 4204–4215.
Pearson,W.R. and Sierk,M.L. (2005) *Curr. Opin. Struct. Biol.*, **15**, 254–260.
Pilone,M.S. (2000) *Cell Mol. Life Sci.*, **57**, 1732–1747.
R Development Core Team. (2003) *R: a language and environment for statistical computing.*
Radivojac,P., Obradovic,Z., Smith,D.K., Zhu,G., Vucetic,S., Brown,C.J., Lawson,J.D. and Dunker,A.K. (2004) *Protein Sci.*, **13**, 71–80.
Rost,B. (1999) *Protein Eng.*, **12**, 85–94.
Ryckaert,J.P., Ciccotti,G. and Berendsen,H.J.C. (1977) *J. Comp. Phys.*, **23**, 327–341.
Sierk,M.L. and Pearson,W.R. (2004) *Prot. Sci.*, **13**, 773–785.
Sing,T., Sander,O., Beerenwinkel,N. and Lengauer,T. (2005) *Bioinformatics*, **21**, 3940–3941.
Smith,D.K., Radivojac,P., Obradovic,Z., Dunker,A.K. and Zhu,G. (2003) *Protein Sci.*, **12**, 1060–1072.
Tai,K. (2004) *Biophys. Chem.*, **107**, 213–220.
Tai,K., Murdock,S., Wu,B., Ng,M.H., Johnston,S., Fangohr,H., Cox,S.J., Jeffreys,P., Essex,J.W. and Sansom,M.S.P. (2004) *Org. Biomol. Chem.*, **2**, 3219–3221.
Todeschini,R. (1997) *Anal. Chim. Acta*, **348**, 419–430.
Todone,F., Vanoni,M.A., Mozzarelli,A., Bolognesi,M., Coda,A., Curti,B. and Mattevi,A. (1997) *Biochemistry*, **36**, 5853–5860.
Vreede,J., van der Horst,M.A., Hellingwerf,K.J., Crielaard,W. and van Aalten,D.M.F. (2003) *J. Biol. Chem.*, **278**, 18434–18439.
Yang,L., Liu,X., Jursa,C.J., Holliman,M., Rader,A.J., Karimi,H.A. and Bahar,I. (2005) *Bioinformatics*, **21**, 2978–2987.
Yue,Q.K., Kass,I.J., Sampson,N.S. and Vrielink,A. (1999) *Biochemistry*, **38**, 4277–4286.
Zheng,W., Brooks,B.R., Doniach,S. and Thirumalai,D. (2005) *Structure*, **13**, 565–577.

Received August 05, 2006; revised April 06, 2007; accepted April 18, 2007

Edited by David Thirumalai