

METHODOLOGY ARTICLE

Open Access

# Dynamics based alignment of proteins: an alternative approach to quantify dynamic similarity

Márton Münz<sup>1,2,3</sup>, Rune Lyngsø<sup>2</sup>, Jotun Hein<sup>2,3</sup>, Philip C Biggin<sup>1,3\*</sup>

## Abstract

**Background:** The dynamic motions of many proteins are central to their function. It therefore follows that the dynamic requirements of a protein are evolutionary constrained. In order to assess and quantify this, one needs to compare the dynamic motions of different proteins. Comparing the dynamics of distinct proteins may also provide insight into how protein motions are modified by variations in sequence and, consequently, by structure. The optimal way of comparing complex molecular motions is, however, far from trivial. The majority of comparative molecular dynamics studies performed to date relied upon prior sequence or structural alignment to define which residues were equivalent in 3-dimensional space.

**Results:** Here we discuss an alternative methodology for comparative molecular dynamics that does not require any prior alignment information. We show it is possible to align proteins based solely on their dynamics and that we can use these dynamics-based alignments to quantify the dynamic similarity of proteins. Our method was tested on 10 representative members of the PDZ domain family.

**Conclusions:** As a result of creating pair-wise dynamics-based alignments of PDZ domains, we have found evolutionarily conserved patterns in their backbone dynamics. The dynamic similarity of PDZ domains is highly correlated with their structural similarity as calculated with Dali. However, significant differences in their dynamics can be detected indicating that sequence has a more refined role to play in protein dynamics than just dictating the overall fold. We suggest that the method should be generally applicable.

## Background

It is well established that conformational flexibility plays a key role in the biochemical functions of proteins [1,2]. Protein motions of functional importance range from fast (sub-nanosecond) atomic fluctuations to slow (microsecond upward), large-scale conformational rearrangements [3,4]. Several studies have managed to relate internal protein motions to biochemical functions [5,6], and in particular the characterization and prediction of large-scale conformational changes via the use of normal modes [7] and elastic-network models [8,9] has been quite successful. However, there are many signalling molecules (for example PDZ domains) where there is not a large-scale conformational change and yet somehow the information that a ligand has bound is

communicated to a different region of the protein. Furthermore, it is not clear if slight variations in structure can lead to large variations in dynamics, or similar protein structures always have similar motions [9]. These problems require a more detailed picture of the underlying dynamics.

Molecular Dynamics (MD) simulations can be used effectively to explore the conformational energy landscape accessible to proteins [10,11] giving an insight into how protein dynamics relates back to sequence. Comparative MD studies address these questions by performing MD simulations of multiple proteins and comparing their dynamic trajectories. Previous comparative MD studies of proteins fall into two main categories. Studies in the first class compared the dynamics of the same protein simulated under different conditions [12-15]. In this case, the question is how the motion of the protein is altered by the new condition, for example

\* Correspondence: philip.biggin@bioch.ox.ac.uk

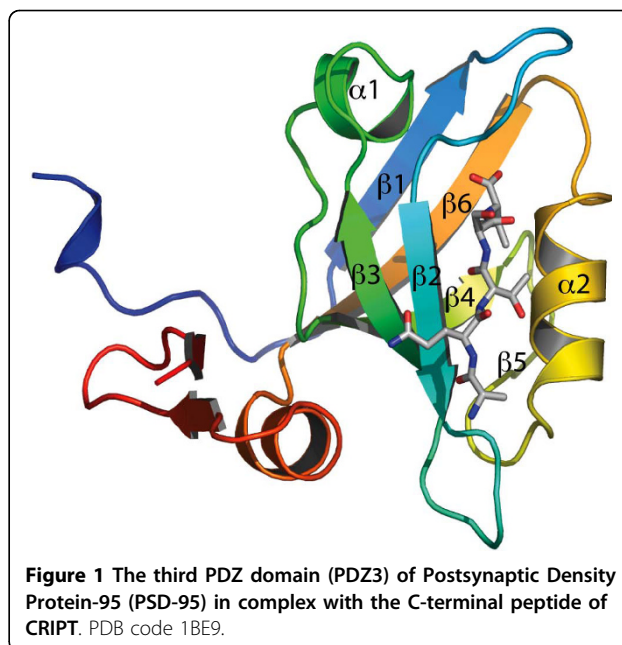
<sup>1</sup>Structural Bioinformatics and Computational Biochemistry, University of Oxford, South Parks Road, Oxford, OX1 3QU, UK

the presence of a ligand. By contrast, the second class of studies compared the dynamics of different proteins simulated under the same condition [16-18] in order to pinpoint similarities and differences in functionally important movements.

If the dynamics of non-identical proteins is compared, a mapping between the different structures is often required. Therefore a common point of previous comparative MD studies of homologous proteins is that they use prior sequence or structure alignments to find residue equivalencies between the proteins. For example, to compare the fluctuation of different cold-active enzymes, Spiwok *et al.* [18] used structural alignment to define equivalent residue pairs between the proteins. Papaleo *et al.* [17] performed MD simulations of different elastases to compare their molecular flexibility. Here the correspondences between residues of different proteins were derived from pairwise sequence alignments. Another example for alignment-guided comparison of protein dynamics is the work of Pang *et al.* [16] who simulated a series of proteins within the same fold family. To compare the fluctuations as well as the principal components of dynamics of only the structurally conserved residues across the set of proteins, they used structural alignment to define conserved positions. Alternatively to MD, simplified models called Gaussian Network Models were used to explore the common dynamics of the globin family [19] and the protease superfamily [20]. In these studies, however, comparative analysis of dynamics also relied on prior alignments of the proteins.

Nevertheless, if dynamically equivalent regions do not match to sequentially and structurally similar regions, alignments can mislead the comparison of protein motions. Recently, Zen *et al.* [21,22] developed a method that takes a combined measure of spatial and dynamic consistency to derive an alignment on the fly that can be used to compare the proteins. We introduce here an alternative method to measure the similarity of backbone dynamics of proteins without the use of any prior alignment information. The method creates pairwise alignments of proteins based solely on their backbone motions without taking into account their sequence and structure. Scores of dynamics-based alignments are used to quantify the dynamic similarity of proteins. The proposed similarity measure can be applied to study the topology of protein dynamics space.

The method is tested on members of the PDZ domain family. PDZ (post-synaptic density-95/discs large/zonula occludens-1) domains are common protein-protein interaction domains [23], most often binding the C-terminal of the ligand protein [24] (see Fig. 1). PDZ domains play an important role in organizing signalling complexes and ion channels [25,26]. The biophysical



aspects of folding and binding reactions of PDZ domains has been intensively studied [27], as well as their ligand preferences [28,29] and the background of binding specificity against a wide range of ligands [30]. Due to their biological importance, PDZ domains are found in a large number of species: bacteria, yeast, plants, invertebrates and vertebrates [31]. The dynamics (collective normal modes) of PDZ domains in relation to functional properties was studied on a coarse-grained model [32]. Recently, Law *et al.* [33] have proven by NMR-relaxation experiments that PDZ domain side-chain dynamics is evolutionary conserved. In this paper we conclude that backbone dynamics of PDZ domains is also evolutionarily conserved. We first describe the development of the method and how it can be used to generate an alignment before demonstrating its application to PDZ domains.

## Results

### Dynamic Fingerprint Matrix

We introduce a novel way to characterize the backbone dynamics of proteins. The underlying idea is that in a moving protein each residue is fluctuating relative to all other residues, therefore a detailed description of motion should capture all inter-residue fluctuations. While static structures are often characterized by the matrix of inter-residue distances (i.e. the distance matrix), this representation is not applicable for a moving protein in which inter-residue distances are constantly changing. It is possible, however, to characterize the relative motion of any two residues by the distribution of their distance over time. We measure the extent

to which residue  $i$  and residue  $j$  are fluctuating relative to each other by

$$F_{ij} = \sigma \{ D_{ij}(t) \} \quad (1)$$

where  $F_{ij}$  is calculated as the standard deviation of the distribution of  $D_{ij}$  (distance of the two residues) in the whole conformational ensemble generated by MD simulation. In this initial investigation we only consider distances between  $C\alpha$  atoms, but the technique can easily be extended to a more detailed description of each amino acid residue. The standard deviation of the distance distribution reflects how much the two residues fluctuate relative to each other.  $F_{ij}$  values are calculated for each residue pair and are collected into a matrix denoted by  $F$ , which we will refer to as the Dynamic Fingerprint Matrix (DFM). Similarly to a distance matrix that characterizes a single conformation, a dynamic fingerprint matrix characterizes an ensemble of conformations. (The relationship of DFMs and correlation matrices is discussed in Additional file 1.)

#### Comparing DFMs using prior alignment

Where a prior alignment is known, the comparison of DFMs is straightforward. Given a pairwise sequence alignment of protein A and B, let  $\alpha$  and  $\beta$  be the index vectors of the aligned residues of sequence A and B, respectively. That is, the  $k^{\text{th}}$  match column, (i.e. columns not containing a gap in the alignment) aligns residue  $\alpha$  ( $k$ ) of protein A with residue  $\beta$  ( $k$ ) of protein B. Thus each pairwise alignment can be characterized by an  $(\alpha, \beta)$  pair. Let  $F^A$  and  $F^B$  be the DFMs of the two proteins. The  $(\alpha, \beta)$  alignment define a submatrix of size of  $|\alpha| \times |\alpha|$  of both DFMs. The  $(i, j)$  entries of the two submatrices are given by

$$F_{\alpha(i)\alpha(j)}^A \quad \text{and} \quad F_{\beta(i)\beta(j)}^B \quad (2)$$

Note that the two submatrices describe the pairwise fluctuations of the aligned residues only. The  $(i, j)$  entries of the two submatrices are considered equivalent as they describe the fluctuations of equivalent pairs of residues.

We define the dynamic similarity score of protein A and B based on a prior  $(\alpha, \beta)$  alignment as:

$$S^{AB}(\alpha, \beta) = \sum_{i=1}^{|\alpha|} \sum_{j=i+1}^{|\alpha|} s(i, j) \quad (3)$$

where each pair of equivalent matrix entries are compared one-by-one and their contribution to the overall score is given by

$$s(i, j) = \frac{s_+ - s_-}{1 + e^{\lambda(d(i,j) - \omega)}} + s_- \quad (4)$$

where

$$d(i, j) = \frac{|F_{\alpha(i)\alpha(j)}^A - F_{\beta(i)\beta(j)}^B|}{(F_{\alpha(i)\alpha(j)}^A + F_{\beta(i)\beta(j)}^B)/2} \quad (5)$$

is the relative difference of the two equivalent matrix entries.

$s(i, j)$  is an S-shaped logistic function that assigns positive score ( $s_+$ ) to highly similar matrix entries and negative score ( $s_-$ ) to highly dissimilar entries (Additional file 1: Fig. S1). An user-adjustable cut-off parameter,  $t$ , defines the critical  $d(i, j)$  over which  $s(i, j)$  turns negative. The relationship between  $t$  and parameter  $\omega$  of  $s(i, j)$  is discussed in Additional file 1 as well as the choice of parameter values. The key difference from using a discrete threshold is that the parameter  $\lambda$  can be tuned to set the steepness of the S-shaped function to make  $s(i, j)$  less dependent on the cut-off parameter  $t$ . Since  $s(i, j)$  is associated with a match column pair in the alignment, it will be referred to as the Pairwise Match Score (PMS) of columns  $i$  and  $j$ .

#### Comparing DFMs without prior alignment

In the previous section we have introduced how to compare the DFMs of two proteins using a prior sequence/structural alignment. Our main goal, however, is to find the optimal alignment of two proteins based on solely their DFMs. Note that it is the opposite strategy of previous comparative MD studies which relied upon prior alignments. We aim to find the  $(\alpha, \beta)$  pair corresponding to the maximal similarity score. Let  $(\alpha^*, \beta^*)$  be the pair of index vectors for which  $S^{AB}(\alpha, \beta)$  is maximal.  $S^{AB}(\alpha^*, \beta^*)$  is then called the dynamic similarity score of protein A and B and is simply denoted by  $S^{AB}$ . The sequence alignment problem is hereby transformed into a matrix alignment problem. Structural alignment algorithms Dali [34] and MatAlign [35] aim to solve the same question when aligning distance matrices. The search space of  $(\alpha, \beta)$  pairs is exponentially large and the global optimization problem is in fact NP-hard. In this case to find the maximum score  $S^{AB}$  we employ a simulated annealing protocol (see Methods).

#### Single Match Score (SMS)

Although the PMS scores concerning a given match column depend on the other match columns in the alignment, it is useful to compare the total contributions of each individual column to the alignment score. For match column  $i$ , the sum of PMS scores with respect to

all other match columns will be referred to as the Single Match Score (SMS):

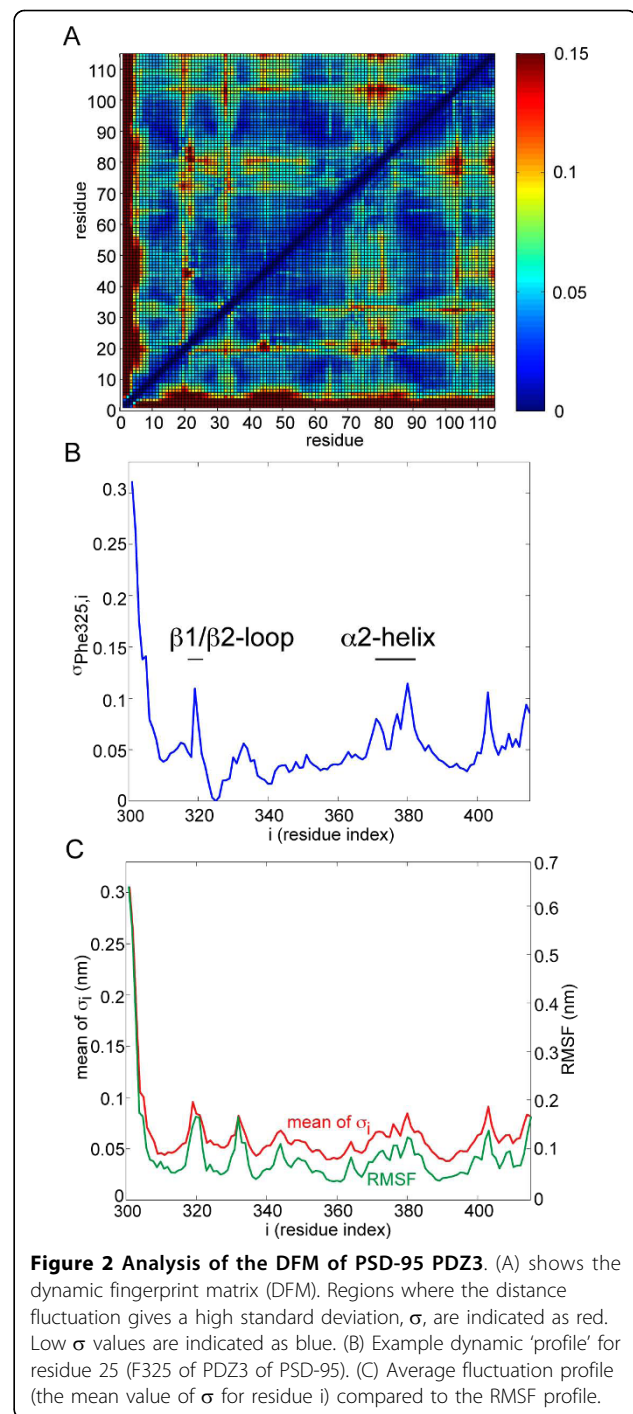
$$s(i) = \sum_{\substack{j=1 \\ j \neq i}}^{|\alpha|} s(i, j) \quad (6)$$

In other words, the SMS of a match column is the score by which the total alignment score decreases in case of removing that match from the alignment. Matches of negative SMS values are beneficial to remove in optimizing the alignment. Hence the optimal dynamics-based alignment contains only positions of non-negative SMS values. Either studying a prior (sequence/structural) or a dynamics-based alignment, the SMS-profile represents our confidence in each aligned pairs of residues.

#### Analysis of the motion of PSD-95 PDZ3

Before we discuss the comparative analysis, we first demonstrate that the DFM protocol is appropriate by characterizing the dynamics of the third PDZ domain (PDZ3) of PSD-95 (Postsynaptic Density Protein 95) from *Rattus norvegicus*. PSD-95 plays an important role in controlling synaptic strength and plasticity in the central nervous system [36]. The 110-residue-long PDZ3 is the most well studied PDZ domain [27] which has a canonical PDZ-domain structure consisting of six  $\beta$ -strands ( $\beta$ 1- $\beta$ 6) and two  $\alpha$ -helices ( $\alpha$ 1 and  $\alpha$ 2). The peptide-binding groove is located between the  $\beta$ 2-strand and  $\alpha$ 2-helix (see Fig. 1). As described in Methods, we used a 20 ns MD trajectory to calculate the Dynamic Fingerprint Matrix (DFM) of PDZ3 (see Fig. 2A). Simple analysis of the DFM revealed that the most dynamic part of the domain is the  $\alpha$ 2 helix (His372-Ala382). This observation is in accordance with De Los Rios *et al* [32] who performed Normal Mode Analysis of a Gaussian Network Model of PDZ3. As the  $\alpha$ 2-helix and the  $\beta$ 2-strand form the binding pocket it seems likely that the considerable relative motion of these two structural components may be related to the capacity to bind ligands.

To see how a particular residue fluctuates relative to all other residues, we examined individual rows of the DFM. We will refer to the  $k^{th}$  row of the DFM as the dynamic profile of residue  $k$ . The dynamic profile of Phe325, for example, shows us that it is the  $\beta$ 1/ $\beta$ 2 loop and the  $\alpha$ 2-helix which fluctuates the most relative to this residue (see Fig. 2B). Phe325 is at the N-terminal end of the  $\beta$ 2-strand, right next to the  $\beta$ 1/ $\beta$ 2 (L1) loop that interacts with the carboxylate terminal of the bound peptide. Therefore, the relative motion of Phe325 and the  $\alpha$ 2-helix at the other side of the binding pocket along with the L1 loop may reflect the manner in which



peptide binding occurs. A further example profile from a residue situated within an alpha helix (Asp 348) is shown Additional file 1.

We have calculated the mean value of each rows of the DFM which we will call the average fluctuation profile (Fig. 2C). Atomic fluctuations are often characterized by the RMSF (root mean square fluctuation). It can be seen that the RMSF profile is very similar to the

average fluctuation profile (Fig. 2C). There was 0.94 correlation between the average fluctuation profile and the RMSF profile. We conclude that a DFM contains the same information as a standard RMSF plot, however, by describing relative inter-residue fluctuations, it gives us a more detailed representation of protein flexibility. More importantly, it does away with the dependency on a single “native” reference structure for defining fluctuations and is simple to compute.

#### Dynamics-based alignments of PDZ domains

We have thus far demonstrated that the DFM methodology can be a useful way to analyse protein motions, but the power of the approach is that it enables us to compare the dynamics of two or more different proteins. Furthermore this information can be used to derive an alignment. To illustrate this we selected 10 PDZ domains from a range of organisms (see Table 1) and ran 20 ns explicit MD simulations (see Methods). We then calculated the DFMs for each protein and created dynamics-based alignments of each pairs of proteins using the matrix alignment algorithm described in Methods. Fig. 3. presents an example: the alignment of PDZ3 of PSD-95 and the PDZ domain of neuronal nitric oxide synthase (nNOS). The alignment does not require any prior sequence or structural information; the two DFMs are the only inputs of the algorithm (Fig. 3A). The most similar submatrix pair found by the algorithm have  $77 \times 77$  entries which are highlighted (in white) in Fig. 3B. The optimal submatrix pair corresponds to a pairwise alignment consisting of 77 aligned residues. Removing the rows and columns of the DFMs that correspond to gaps in the alignment, the remaining matrices will be referred to as the ‘collapsed DFMs’ (which are identical to the submatrices identified). Although one cannot see notable similarity between the original DFMs, the collapsed DFMs appear to be visually similar patterns (Fig. 3C).

The derived dynamics-based alignment was compared to a structural alignment created by pairwise DaliLite [37] and a pairwise sequence alignment created by the

Needleman-Wunsch algorithm [38] using EMBOSS-Align [39]. Fig. 4. presents the three alignments annotated by the secondary structure elements of the canonical PDZ-domain fold (i.e. six beta-strands,  $\beta 1$  to  $\beta 6$ , and two alpha-helices,  $\alpha 1$  and  $\alpha 2$ ). For the DFM-alignment and the sequence alignment, the SMS score of each column is also presented, reflecting our confidence in individual aligned positions.

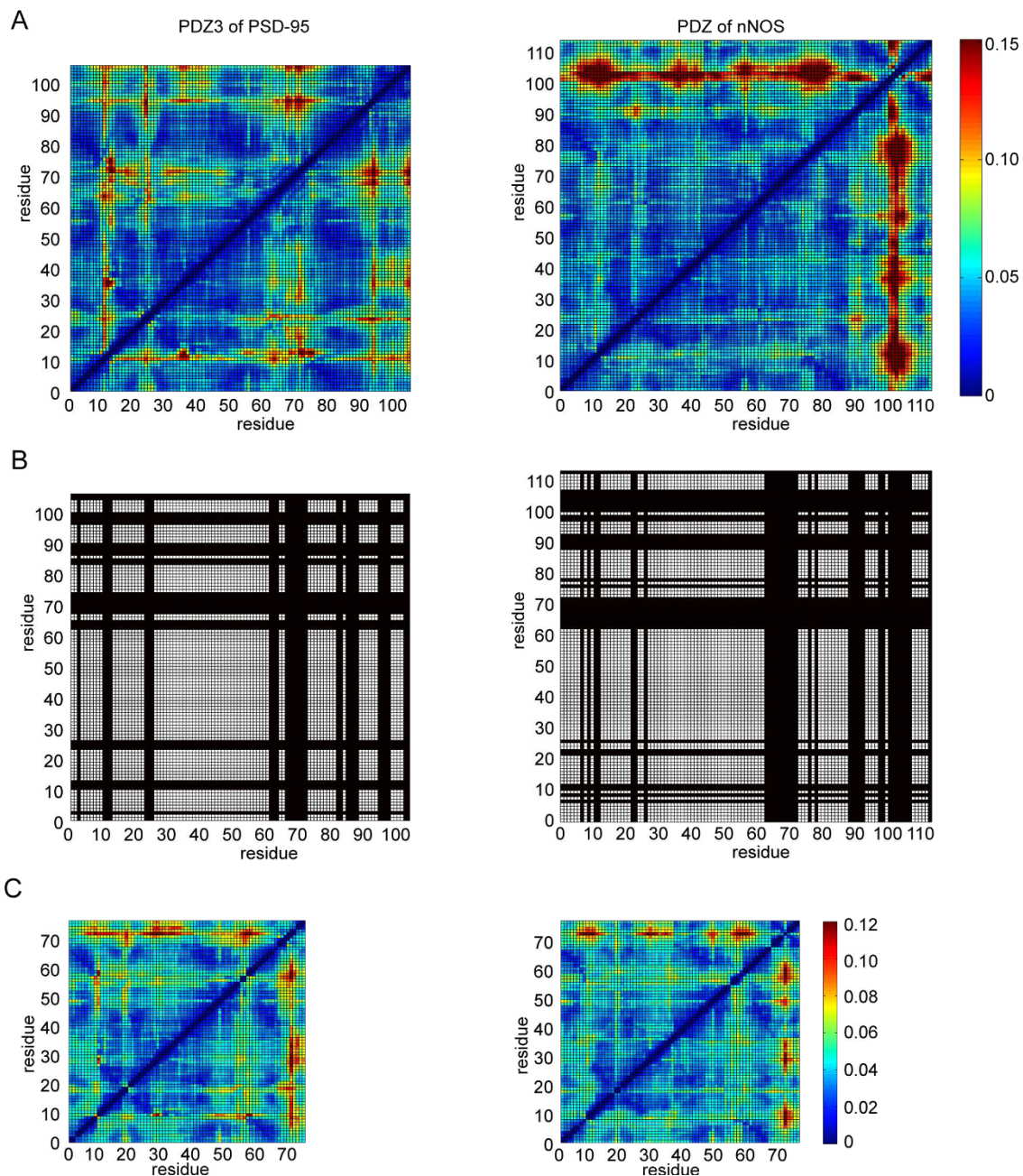
As shown by Fig 4, equivalent secondary structure elements of the two proteins align very well in the dynamics-based alignment, suggesting that dynamics, just like sequence and structure, may contain enough information to match proteins at the secondary structure level. Moreover, the DFM-based alignment includes 20 pairs of identical residues, out of which 18 and 20 are also present in the Needleman-Wunsch and Dali alignments, respectively. Despite all these similarities, however, a striking difference can be seen between the DFM-based alignment and the sequence/structural alignments. The second alpha-helix ( $\alpha 2$ ), included both in the Needleman-Wunsch and Dali alignments, is almost completely missing from the DFM-alignment, indicating that, although conserved at the sequence and structure level, this helix has different dynamics in the two proteins. Characterized above, the  $\alpha 2$ -helix has high mobility in the PDZ3 of PSD-95 unlike in the PDZ of nNOS, that makes the two regions dynamically non-alignable. This is a clear example, when the dynamics-based alignment gives similar information as sequence and structural alignments, but at the same time, it provides new insights into the properties of proteins, that cannot be detected through standard alignment methods.

#### Analysing SMS-profiles

Since it was optimized by the matrix alignment algorithm, the DFM-alignment includes only matches of positive SMS values as shown by Fig. 4A. The SMS-profile has its peaks within  $\beta$ -strands but drops at certain match columns (e.g. in the  $\beta 1/\beta 2$  and  $\beta 2/\beta 3$  loops and

**Table 1 Representative members of the PDZ domain family used in this study.**

Protein containing PDZ domain	PDB entry	Resolution (Å)	Source organism
nNOS	1qau	1.25	<i>Rattus norvegicus</i>
InaD	1ihj	1.8	<i>Drosophila melanogaster</i>
PSD-95 (PDZ3)	1bfe/1be9	2.3/1.82	<i>Rattus norvegicus</i>
tricorn protease	1k32	2.0	<i>Thermoplasma acidophilum</i>
GRIP2 (PDZ4)	1x5r	NMR structure	<i>Homo sapiens</i>
Rv0983	1y8t	2.0	<i>Mycobacterium tuberculosis h37rv</i>
PhotosystemII D1 Protease	1fc6	1.8	<i>Scenedesmus obliquus</i>
Alpha-1 Syntrophin	1qav	1.9	<i>Mus musculus</i>
EpsC	2i6v	1.63	<i>Vibrio cholerae</i>
DVL2	2f0a	1.8	<i>Xenopus laevis</i>

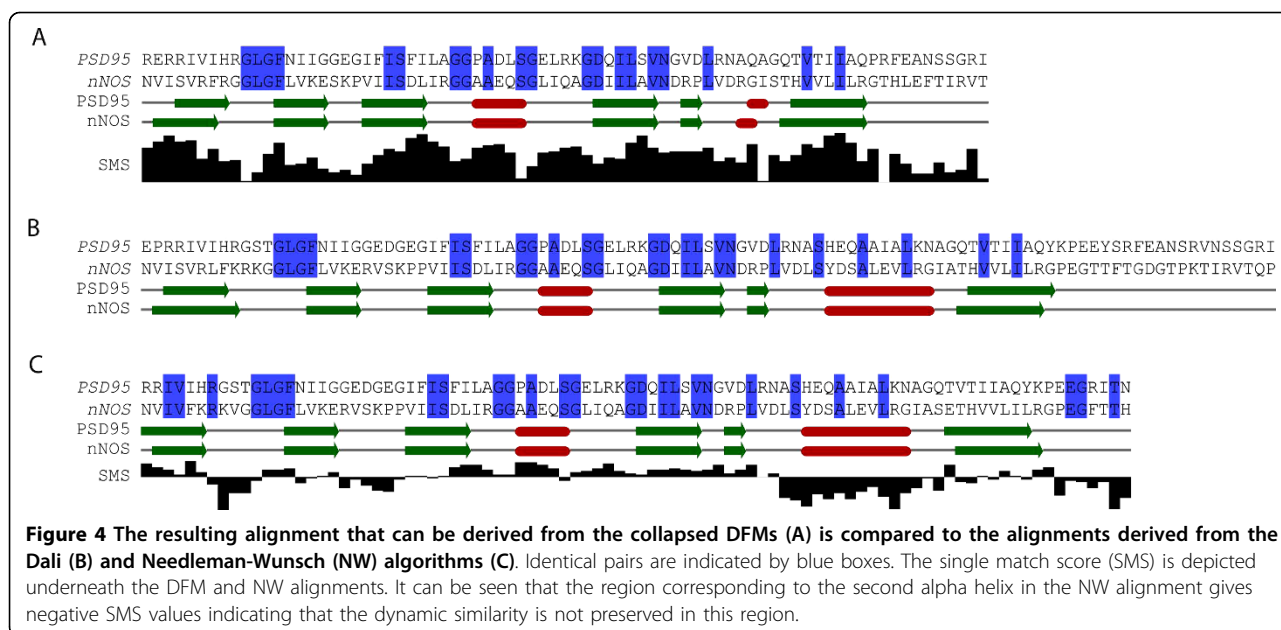


**Figure 3** (A) Comparison of DFMs for PDZ3 of PSD-95 with the PDZ domain of nNOS. (B) Identification of similar submatrices, containing 77 residues in this case, from the DFMs. (C) Collapsed DFMs highlighting the similarities in these proteins.

at the C-terminal end of  $\alpha$ 1-helix). This suggests that  $\beta$ -strands of the domains have minor fluctuations that makes them easier to align than the other regions of the proteins.

As we discussed before, the dynamic similarity of proteins can also be measured based on a prior (sequence or structural) alignment. In this case, the motion of the subsets of residues defined by the prior alignment is compared. To test this option, we used the Needleman-

Wunsch alignment as a prior alignment, which resulted in a dynamic similarity score of -132.8. The optimal similarity score found for this example is 1307.3, and the extreme non-optimality of the sequence based alignment score illustrates that conserved sequence positions can match dynamically dissimilar subsets of residues. Accordingly, 46 per cent of the columns of the collapsed Needleman-Wunsch alignment have negative SMS. The less-matching region (a continuous block of negative



SMS values) appears to be the  $\alpha 2$ -helix, explaining why this region is excluded from the optimized DFM-alignment.

#### Dynamics-space of PDZ domains

The dynamic similarity score of the PDZ domain of nNOS and PDZ3 of PSD-95 is  $S = 1307.3$ , which was converted to a  $p$ -value of  $9 \cdot 10^{-11}$  using the significance analysis framework described in Methods. Likewise the alignment algorithm has found significant dynamic similarities between other pairs of PDZ domains. The  $p$ -values are summarized in Table 2 and can be shown as a dynamic similarity graph Additional file 1: Fig S4A, in which the different proteins are represented by the nodes of the graph, and two proteins are connected if they have significantly similar ( $p$ -value  $< 0.05$ ) dynamics.

It is interesting to note that the dynamic similarity shows differences between different structures. We can see a cluster of five proteins (1be9, 1qau, 1qav, 1ihj and 2f0a) that are better-connected in the dynamic similarity graph. Out of the 10 possible links between these structures 8 are present in the graph. Two additional links are found between 2f0a/1x5r and 1ihj/1fc6. Three structures (2i6v, 1k32 and 1y8t), however, do not have significant dynamic similarity with any other structures. Looking for structural explanations for these differences, we examined the pairwise Dali Z-scores between the 10 domains (summarized in Additional file 1: Table S1 and shown as a Dali Z-score graph in Additional file 1: Fig. S4B). In this second graph, two nodes are connected if their Dali Z-score is more than 8.5 (a threshold selected empirically). As expected, each protein pair has

significant structural similarity (all Dali Z-scores are over 3.5), but a subset of structures are more similar to each other than to the others. A cluster of six structures (1be9, 1qau, 1qav, 1ihj, 2f0a and 1x5r) appears to be fully connected in the graph, while two additional links are found between 1fc6/1y8t and 1ihj/1fc6. Two domains (1k32 and 2i6v) are not linked to any other structures.

The almost perfect overlap (with the only exception of 1x5r) between the well-connected clusters in the two graphs suggests a topology-preserving mapping between the structure space and dynamics space of PDZ domains. There appears to be a strong correlation (0.82) between the raw dynamic similarity scores and Dali Z-scores considering all 45 protein pairs (Additional file 1: Fig. S5), and a weaker but still strong correlation (0.63)

**Table 2 Dynamic similarity  $p$  values for the 10 PDZ domains.**

	1qau	1ihj	1be9	1k32	1x5r	1y8t	1fc6	1qav	2i6v	2f0a
1qau										
1ihj	<u><math>4 \cdot 10^{-4}</math></u>									
1be9	<u><math>9 \cdot 10^{-11}</math></u>	0.34								
1k32	0.62	0.76	0.74							
1x5r	0.32	0.32	0.60	0.39						
1y8t	0.61	0.47	0.58	0.62	0.36					
1fc6	0.08	<u><math>8 \cdot 10^{-3}</math></u>	0.78	0.76	0.65	0.19				
1qav	<u><math>1 \cdot 10^{-8}</math></u>	<u><math>2 \cdot 10^{-3}</math></u>	<u><math>1 \cdot 10^{-5}</math></u>	0.66	0.14	0.07	0.17			
2i6v	0.68	0.78	0.58	0.70	0.71	0.55	0.81	0.59		
2f0a	<u><math>1 \cdot 10^{-4}</math></u>	<u><math>3 \cdot 10^{-6}</math></u>	0.27	0.30	<u><math>8 \cdot 10^{-4}</math></u>	0.14	0.47	<u><math>6 \cdot 10^{-6}</math></u>	0.72	

Those that are significant (at the  $p < 0.05$ ) are highlighted.

considering only the 35 protein pairs having non-significant dynamic similarities.

Interestingly, all the 6 proteins in the fully connected cluster of the Dali graph (5 of which are well-connected in the dynamics graph too) are from multicellular organisms (metazoa), while the other 4 proteins are from unicellular species. The structural difference between PDZ domains from simple and complex organisms is well-known. First recognized by *Liao et al.* [40] and exemplified by other authors [41,42], PDZ domains of bacterial and plant origin have a circularly permuted fold compared to the canonical PDZ domain fold found in metazoa. Despite their considerably different architecture, non-metazoan PDZ domains have very similar overall tertiary structure to metazoan PDZ domains. This is indeed reflected by Additional file 1: Table S1 which shows that metazoan and non-metazoan PDZ domains are significantly similar structures (Dali Z-scores above 2). On the other hand, the fact that the metazoan structures form a distinct cluster in the Dali graph shows the difference between the canonical and the circularly permuted fold. Putting it all together, our data suggests that the essential structural difference between PDZ domains of metazoan and non-metazoan origin is also reflected by the dissimilarity of their dynamics. Metazoan PDZ domains appear to be structurally and dynamically more conserved. However, even within the cluster of the metazoan proteins, there are significant differences in dynamics (on this timescale) that can be quantified.

### Robustness of DFMs

MD simulations are subject to sampling problems. In order to assess whether the simulations have run long enough to provide a reasonable picture of the dynamics we examined the convergence of the DFM patterns. We ran five 20 ns simulations of the same protein (PSD-95 PDZ3) using different random seeds for the initial atomic velocities. The similarities of each pairs of DFMs resulted from the different MD runs were measured by the matrix alignment algorithm. Naturally, the five DFMs were not perfectly the same, but the similarity between each pair was highly significant (see Table 3). Most importantly, comparing different simulations of the same protein results in much higher similarity scores, than the comparison of different PDZ domains. These results lead us to the conclusion that the sampling in 20 ns simulations can be sufficient to provide robust DFM patterns for a comparative analysis. Clearly, simulating the proteins for longer period of time further improves the convergence of DFMs.

### Conclusions

We have demonstrated a novel methodology for comparing the backbone dynamics of proteins simulated by

**Table 3 Dynamic similarity scores of five trajectories of 1be9 with different initial seed velocities.**

	Run 1	Run 2	Run 3	Run 4	Run 5
Run 1					
Run 2	<u>2363.3</u>				
Run 3	<u>3223.7</u>	<u>3029.8</u>			
Run 4	<u>2283.9</u>	<u>1589.4</u>	<u>1715.7</u>		
Run 5	<u>2693.1</u>	<u>2221.1</u>	<u>2692.8</u>	<u>2244.1</u>	

All pairs are highly significant (p-value = 0).

Molecular Dynamics simulations and for deriving an alignment that is based solely on the underlying dynamics of the system within a particular timescale; in this case 20 ns. We have selected that timescale for practical reasons, but it is effectively arbitrary as we were mainly interested in seeing if we could quantify the similarity in a meaningful way. The method should work just as well for longer timescales. Of course, the most useful timescale is the one which proteins exhibit their function. The comparison of our dynamic similarity score with the Dali score demonstrates structure and dynamics are indeed correlated, but at a level ( $R = 0.82$ ) that still allows for significant differences in dynamics to be apparent. This may explain why methods such as elastic network models seem to work for predicting large-scale conformational changes, but that the detailed differences between protein motions can still be significant. With respect to the PDZ domains we have shown that the majority of motions within the protein are similar, but additionally that we can detect differences (located at the binding interface) that are significantly different and could not be detected by the usual sequence or structure based alignment methods. We have developed a method for measuring dynamic similarity between proteins with a simple algorithm. As the method is capable of detecting precise differences in the dynamics between structures it could also be used to assess the influence of ligand-binding on the dynamics of protein structure. We are currently exploring that aspect as well as developing further improvements on the algorithm.

### Methods

#### Molecular Dynamics Simulations

We carried out all-atom MD simulations of 10 representative members of the PDZ domain family (Table 1) in explicit water and with  $\text{Na}^+$  and  $\text{Cl}^-$  ions to a concentration of 150 mM. After 200 ps of restrained MD, 20 ns of unrestrained MD was performed with the GROMACS software package [43,44] using the OPLS force field [45] in an NPT ensemble (see Additional file 1 for full details). Snapshots from the trajectories were saved every 5 ps for analysis.



### Matrix Alignment Algorithm

To find a good approximation for the global maximum of  $S^{AB}(\alpha, \beta)$  in a reasonable time, we have developed a heuristics approach based on the multiple restart Simulated Annealing (SA) method. The algorithm performs an MCMC (Markov chain Monte Carlo) search in the space of  $(\alpha, \beta)$  pairs. The Markov chain starts from a random initial alignment, and in each step the alignment is modified by inserting or removing one residue pair. We use the Metropolis acceptance criterion [46] to decide the next state of the chain. The parameter called 'temperature' which controls the acceptance probability is gradually reduced according to an exponential decay annealing schedule, leading to the convergence to a high-scoring and potentially optimal alignment. We let the Markov chain explore the search space at a given constant temperature: the chain has to go through a minimal number of accepted steps before the temperature is further reduced. The initial temperature is calibrated using the method proposed by Johnson et al. [47]. The whole SA procedure terminates when the acceptance ratio goes below a critical value. To overcome the stochastic nature of SA and the possible existence of local optima, the process is restarted for a number of times from random initial states and the best result of the multiple runs is selected as the final output of the algorithm.

### Significance Analysis

To assess the statistical significance of dynamic similarity scores between PDZ domains, we have performed 20 ns MD simulations of 12 evolutionarily and functionally unrelated proteins of different sizes referred to as the Reference Set (Additional file 1: Table S2). Reference proteins were aligned using our dynamic fingerprint alignment algorithm to measure the background distribution of similarity scores. Using the background distribution, the significance of dynamic similarity of any two proteins can be expressed by the p-value of their similarity score.

Since the optimal alignment score of two proteins is the maximum of the scores of their possible alignments, it follows a type I Extreme Value Distribution, as shown in Additional file 1: Fig. S2. The background score distribution was found to be dependent on the lengths of sequences (i.e. the size of input DFMs). To capture the size dependency of the background distribution, we approximated the  $\mu(L)$  and  $\sigma(L)$  functions, where  $\mu$  is the location parameter,  $\sigma$  is the scale parameter of the Extreme Value Distribution and  $L = L_A L_B$  is the product of the two sequence lengths.

To measure  $\mu(L)$  and  $\sigma(L)$  in a wider interval, not only the full DFMs but their submatrices of different sizes were also aligned. A total number of 2970 dynamic alignments were performed. Regression lines were fitted

to the measured points of  $\mu(L)$  and  $\sigma(L)$  and are used to approximate the parameters of the background distribution (Additional file 1: Fig. S3). The p-value of a dynamic alignment of protein A and B is therefore given by

$$p(A, B) = \exp \left[ -\exp \left( \frac{S^{AB} - \mu(L)}{\sigma(L)} \right) \right] \quad (7)$$

where  $L = L_A L_B$ , and  $L_A$  and  $L_B$  are the lengths of protein A and B, respectively. (For a more detailed explanation of the applied significance analysis framework, see Additional file 1.) We use a significance threshold of  $p = 0.05$  to detect significant dynamic similarities between proteins.

**Additional file 1: Supplementary Information.** Contains detailed information about the actual algorithms employed, data sets used, and extended interpretation of the data presented in the main manuscript.

### Abbreviations

PDZ: post-synaptic density-95/discs large/zonula occludens-1; DFM: dynamics fingerprint matrix.

### Acknowledgements

We thank the Wellcome Trust, the John Fell Fund, Polygene (LSHC-CT-2005-018827), Tom Darden, the NIH and the Systems Biology Doctoral Training Centre for support. PCB is an RCUK Fellow. We thank the anonymous reviewers for helpful suggestions.

### Author details

<sup>1</sup>Structural Bioinformatics and Computational Biochemistry, University of Oxford, South Parks Road, Oxford, OX1 3QU, UK. <sup>2</sup>Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG, UK. <sup>3</sup>Oxford Centre for Integrative Systems Biology, Department of Biochemistry, South Parks Road, Oxford, OX1 3QU, UK.

### Authors' contributions

JH and PCB designed the research. MM performed the research. MM, RL and PCB analyzed the data. MM, RL and PCB wrote the paper. All authors read and approved the final manuscript.

Received: 11 November 2009 Accepted: 14 April 2010

Published: 14 April 2010

### References

1. Dodson G, Verma CS: Protein flexibility: Its role in structure and mechanism revealed by molecular simulations. *Cellular and Molecular Life Sciences* 2006, **63**(2):207-219.
2. Teilum K, Olsen JG, Kragelund BB: Functional aspects of protein flexibility. *Cellular and Molecular Life Sciences* 2009, 1-17.
3. Boehr DD, Dyson HJ, Wright PE: An NMR perspective on enzyme dynamics. *Chemical Reviews* 2006, **106**(8):3055-3079.
4. Henzler-Wildman KA, Lei M, Thai V, Kerns SJ, Karplus M, Kern D: A hierarchy of timescales in protein dynamics is linked to enzyme catalysis. *Nature* 2007, **450**(7171):913-916.
5. Hammes-Schiffer S, Benkovic SJ: Relating protein motion to catalysis. *Annual Review of Biochemistry* 2006, **75**:519-541.
6. Smock RG, Gierasch LM: Sending signals dynamically. *Science* 2009, **324**(5924):198-203.
7. Ma J: Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. *Structure* 2005, **13**:373-380.

8. Hall BA, Kaye SL, Pang A, Perera R, Biggin PC: **Characterization of Protein Conformational States by Normal-Mode Frequencies.** *J Am Chem Soc* 2007, **129**(37):11394-11401.
9. Keskin O, Jernigan RL, Bahar I: **Proteins with similar architecture exhibit similar large-scale dynamic behaviour.** *Biophys J* 2000, **78**:2093-2106.
10. Karplus M, Kuriyan J: **Molecular dynamics and protein function.** *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(19):6679-6685.
11. Karplus M, McCammon JA: **Molecular dynamics simulations of biomolecules.** *Nat Struct Biol* 2002, **9**:646-652.
12. Cox K, Sansom MSP: **One membrane protein, two structures and six environments: A comparative molecular dynamics simulation study of the bacterial outer membrane protein PagP.** *Molecular Membrane Biology* 2009, **26**(4):205-214.
13. Lama D, Sankaramakrishnan R: **Anti-apoptotic Bcl-XL protein in complex with BH3 peptides of pro-apoptotic Bak, Bad, and Bim proteins: Comparative molecular dynamics simulations.** *Proteins: Structure, Function and Genetics* 2008, **73**(2):492-514.
14. Yaneva R, Springer S, Zacharias M: **Flexibility of the MHC class II peptide binding cleft in the bound, partially filled, and empty states: A molecular dynamics simulation study.** *Biopolymers* 2009, **91**(1):14-27.
15. Zacharias M, Springer S: **Conformational flexibility of the MHC Class alpha1-alpha2 domain in peptide bound and free states: A molecular dynamics simulation study.** *Biophysical Journal* 2004, **87**(4):2203-2214.
16. Pang A, Arinaminpathy Y, Sansom MSP, Biggin PC: **Comparative Molecular Dynamics: Similar folds and similar motions?** *Proteins: Struct Funct Bioinf* 2005, **61**:809-822.
17. Papaleo E, Riccardi L, Villa C, Fantucci P, De Gioia L: **Flexibility and enzymatic cold-adaptation: A comparative molecular dynamics investigation of the elastase family.** *Biochimica et Biophysica Acta - Proteins and Proteomics* 2006, **1764**(8):1397-1406.
18. Spiwok V, Lipovová P, Skálová T, Dusková J, Dohnálek J, Hasek J, Russell NJ, Králová B: **Cold-active enzymes studied by comparative molecular dynamics simulation.** *Journal of Molecular Modeling* 2007, **13**(4):485-497.
19. Maguid S, Fernandez-Alberti S, Ferrelli L, Echave J: **Exploring the common dynamics of homologous proteins. Application to the globin family.** *Biophysical Journal* 2005, **89**(1):3-13.
20. Carnevale V, Raugei S, Micheletti C, Carloni P: **Convergent dynamics in the protease enzymatic superfamily.** *J Am Chem Soc* 2006, **128**:9766-9772.
21. Zen A, Carnevale V, Lesk AM, Micheletti C: **Correspondences between low-energy modes in enzymes: dynamics-based alignment of enzymatic functional families.** *Protein Sci* 2008, **17**(5):918-929.
22. Zen A, de Chiara C, Pastore A, Micheletti C: **Using dynamics-based comparisons to predict nucleic acid binding sites in proteins: an application to OB-fold domains.** *Bioinformatics* 2009, **25**(15):1876-1883.
23. Fanning AS, Anderson JM: **PDZ domains: Fundamental building blocks in the organization of protein complexes at the plasma membrane.** *Journal of Clinical Investigation* 1999, **103**(6):767-772.
24. Van Ham M, Hendriks W: **PDZ domains - Glue and guide.** *Molecular Biology Reports* 2003, **30**(2):69-82.
25. Harris BZ, Lim WA: **Mechanism and role of PDZ domains in signaling complex assembly.** *Journal of Cell Science* 2001, **114**(18):3219-3231.
26. Jelen F, Oleksy A, Smietana K, Otlewski J: **PDZ domains - Common players in the cell signaling.** *Acta Biochimica Polonica* 2003, **50**(4):985-1017.
27. Jemth P, Gianni S: **PDZ Domains: Folding and Binding.** *Biochemistry* 2007, **46**(30):8701-8708.
28. Hung AY, Sheng M: **PDZ domains: Structural modules for protein complex assembly.** *Journal of Biological Chemistry* 2002, **277**(8):5699-5702.
29. Tonikian R, Zhang Y, Sazinsky SL, Currell B, Yeh J-H, Reva B, Held HA, Appleton BA, Evangelista M, Wu Y, et al: **A Specificity Map for the PDZ Domain Family.** *PLoS Biol* 2008, **6**(9):e239.
30. Chen JR, Chang BH, Allen JE, Stiffler MA, MacBeath G: **Predicting PDZ domain-peptide interactions from primary sequences.** *Nature Biotechnology* 2008, **26**(9):1041-1045.
31. Ponting CP: **Evidence for PDZ domains in bacteria, yeast, and plants.** *Protein Science* 1997, **6**(2):464-468.
32. De Los Rios P, Cecconi F, Pretre A, Dietler G, Michielin O, Piazza F, Juanico B: **Functional Dynamics of PDZ Binding Domains: A Normal-Mode Analysis.** 2005, **89**(1):14-21.
33. Law AB, Fuentes EJ, Lee AL: **Conservation of side-chain dynamics within a protein family.** *J Am Chem Soc* 2009, **131**(18):6322-6323.
34. Holm L, Sander C: **Protein structure comparison by alignment of distance matrices.** *Journal of Molecular Biology* 1993, **233**(1):123-138.
35. Aung A, Tan K-L: **MatAlign: Precise protein structure comparison by matrix alignment.** *J Bioinf Comp Biol* 2006, **4**:1197-1216.
36. Beique JC, Andrade R: **PSD-95 regulates synaptic transmission and plasticity in rat cerebral cortex.** *Journal of Physiology* 2003, **546**(3):859-867.
37. Holm L, Park J: **DaliLite workbench for protein structure comparison.** *Bioinformatics* 2000, **16**(6):566-567.
38. Needleman SB, Wunsch CD: **A general method applicable to the search for similarities in the amino acid sequence of two proteins.** *J Mol Biol* 1970, **48**(3):443-453.
39. Rice P, Longden I, Bleasby A: **EMBOSS: The European Molecular Biology Open Software Suite.** *Trends in Genetics* 2000, **16**(6):276-277.
40. Liao D-I, Qian J, Chisholm DA, Jordan DB, Diner BA: **Crystal structures of the photosystem II D1 C-terminal processing protease.** *Nat Struct Bio* 2000, **7**:749-753.
41. Krojer T, Garrido-Franco M, Huber R, Ehrmann M, Clausen T: **Crystal structure of DegP (HtrA) reveals a new protease-chaperone machine.** *Nature* 2002, **416**(6879):455-459.
42. Wilken C, Kitzing K, Kurzbauer R, Ehrmann M, Clausen T: **Crystal Structure of the DegS Stress Sensor: How a PDZ Domain Recognizes Misfolded Protein and Activates a Protease.** *Cell* 2004, **117**(4):483-494.
43. Lindahl E, Hess B, Spoel van der D: **GROMACS 3.0: A package for molecular simulation and trajectory analysis.** *J Mol Model* 2001, **7**:306-317.
44. Spoel van der D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC: **GROMACS: fast, flexible and free.** *J Comput Chem* 2005, **26**:1701-1718.
45. Jorgensen WL, Tirado-Rives J: **The OPLS potential functions for proteins. Energy minimizations for crystals of cyclic peptides and crambin.** *J Am Chem Soc* 1988, **110**:1657-1666.
46. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E: **Equation of state calculations by fast computing machines.** *J Chem Phys* 1953, **21**:1087-1092.
47. Johnson DS, Aragon CR, McGeoch LA, Schevon C: **Optimization by Simulated Annealing: An Experimental Evaluation; Part I, Graph Partitioning.** *Operations Research* 1989, **37**(6):865-892.

doi:10.1186/1471-2105-11-188

**Cite this article as:** Münz et al.: Dynamics based alignment of proteins: an alternative approach to quantify dynamic similarity. *BMC Bioinformatics* 2010 **11**:188.

**Submit your next manuscript to BioMed Central and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
www.biomedcentral.com/submit

