W246–W251 *Nucleic Acids Research, 2008, Vol. 36, Web Server issue* *Published online 14 May 2008*
doi:10.1093/nar/gkn259

pFlexAna: detecting conformational changes in remotely related proteins

Anshul Nigham[1,2], Lisa Tucker-Kellogg[1,2], Ivana Mihalek[3], Chandra Verma[3] and David Hsu[1,4,*]

[1]Department of Computer Science, National University of Singapore, Singapore 117590, [2]Singapore–MIT Alliance, Singapore 117576, [3]Bioinformatics Institute (A*STAR), Singapore 138671 and [4]Graduate School of Integrative Sciences & Engineering, National University of Singapore, Singapore 117456

Received February 21, 2008; Revised April 11, 2008; Accepted April 20, 2008

ABSTRACT

The pFlexAna (protein flexibility analyzer) web server detects and displays conformational changes in remotely related proteins, without relying on sequence homology. To do so, it first applies a reliable statistical test to align core protein fragments that are structurally similar and then clusters these aligned fragment pairs into 'super-alignments', according to the similarity of geometric transformations that align them. The result is that the dominant conformational changes occur between the clusters, while the smaller conformational changes occur within a cluster. pFlexAna is available at http://bigbird.comp.nus.edu.sg/pfa2/.

INTRODUCTION

Conformational change plays a critical role in the functioning and regulation of many proteins, and comparing protein structures with different backbone conformations is a common task in structural biology (1). This task is particularly challenging when we compare two evolutionarily divergent proteins. The main goal of our work is to provide an automated tool for detecting conformational changes in remotely related proteins.

For proteins undergoing conformational change, we can often find in them backbone fragments that remain rigid and are well aligned. However, these rigid local fragments reorient themselves with respect to one another during the conformational change, resulting in poor global alignment. A good example is large-scale domain movement. Consider the HECT domain of the human ubiquitin ligase WWP1, which is homologous to the human ubiquitin ligase E6AP. Comparing their crystal structures [PDB codes 1D5F:C (2) and 1ND7 (3)] reveals a dramatic conformational change (Figure 1) crucial for their biological function, which is to move the ubiquitin molecule from one substrate protein to another. The crystallographers who determined the WWP1 structure described in detail its comparison with E6AP (3), and they produced several figures depicting superpositions carried out manually, similar to what pFlexAna produces automatically. To our knowledge, there are currently no other computational tools that aim to automate such analyses.

Our pFlexAna (protein flexibility analyzer) web server aims at automatically detecting conformational changes between a pair of protein structures, without relying on sequence homology. Specifically, it detects regions exhibiting structural change, contrasted with regions exhibiting structural similarity. The pFlexAna results can be viewed as indications of the endpoints of hypothesized molecular motions or mutation-induced conformational changes.

To detect conformational changes in dissimilar proteins, we must first find structurally similar fragments (domains, sub-domains, etc.) in the proteins. pFlexAna employs a statistical test to match core protein fragments that are structurally similar and determine the geometric transformations that align them. This statistical test of structural similarity has been shown to be effective in detecting many different types of conformational changes (4). pFlexAna displays the protein structures superimposed according to each aligned fragment pair. The superimposed display highlights the discontinuity from the similar to the dissimilar regions of the proteins and aids visualization of the conformational changes that occur relative to the aligned fragment pair used as the basis for superposition. See Figure 3 for a sample result. Furthermore, pFlexAna clusters the aligned fragment pairs into 'super-alignments', according to the similarity of geometric transformations that align them. The result is that the dominant conformational changes occur between the clusters, while the smaller conformational changes occur within a cluster.

There are many approaches for studying conformational changes. One possibility is to perform molecular dynamics (MD) simulation to generate alternative conformations from a single known protein structure (5). However, MD simulation is computationally intensive,

*To whom correspondence should be addressed. Tel: +65 6516 2978; Fax: +65 6779 4580; Email: dyhsu@comp.nus.edu.sg

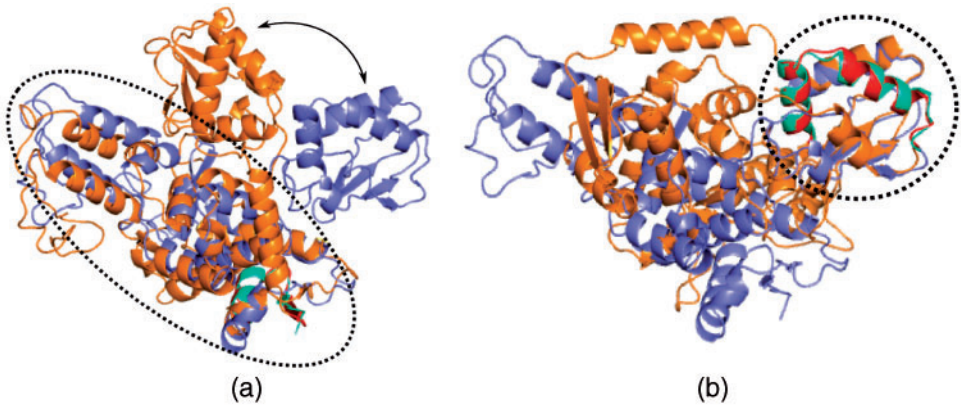

Figure 1. Superposition of the structures of ubiquitin ligases WWWP1 and E6AP. The two different superpositions, produced automatically by pFlexAna, are obtained by aligning a pair of fragments (marked in red and green) with high structural similarity. Dotted ovals annotate the parts of the proteins that are well aligned in the superpositions. To see the magnitude of the conformational change, note that the pose of E6AP (blue) remains the same in both (**a**) and (**b**). Only the pose of WWWP1 (orange) changes.

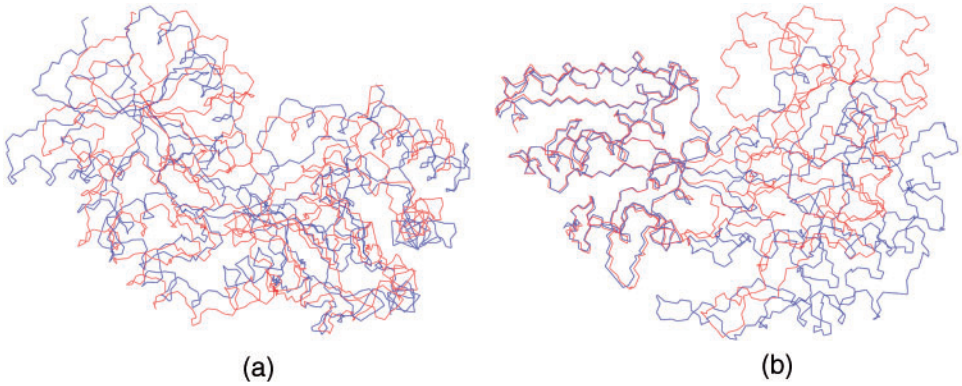

Figure 2. Superposition of the bound (red) and unbound (blue) structures of the maltose-binding protein. (**a**) FATCAT aligns the entire protein structures. (**b**) pFlexAna correctly identifies two rigid domains that undergo hinge bending. The proteins are shown aligned on one domain.

and usually it can only explore conformational changes that are small in magnitude. An alternative approach is to compare directly the structures of a protein in different conformations. There are several good methods for this (4,6–8), but they can be applied only if the structures of the same protein in different conformations are available.

It is thus sometimes necessary to compare conformations of related, but different proteins in order to infer conformational changes. One possibility is to perform sequence alignment as a preprocessing step to match the proteins and then detect conformational changes between them (9). This method, however, is restricted to proteins with 90% sequence identity (9).

As a complementary approach, for cases with low sequence homology, one may attempt structural alignment i.e. matching the proteins based on structural similarity instead of sequence similarity. However, both rigid and flexible structural alignment methods focus on finding a single best fit between the protein structures, and may ignore small yet significant conformational changes. Consider, for example, the relatively simple case of a same protein in two conformations. Given the bound and unbound structures [PDB codes 3MBP (10) and 1LLS:A (11)] of the maltose-binding protein, two popular structural alignment methods, CE (12) and FATCAT (13), align the entire structures together without recognizing the hinge-like conformational change occurring between the two domains of the protein. FlexProt (14), on the other hand, requires the user to know the number of hinges in order to interpret its results. What we need in this case is a method that automatically detects the 'discontinuity' in structural similarity at the hinge due to the conformational change and matches the two domains on each side of the hinge separately. See Figure 2 for a comparison between the results from FATCAT and pFlexAna. In general, pFlexAna aims to detect such conformational changes for two related, but different proteins. Furthermore, it tries to identify and group together those protein fragments that move in tandem.

METHODS

Input

The pFlexAna server takes as input two protein structures in PDB format, which may be uploaded or specified using PDB codes. The user may also specify chain identifiers or a restricted range of residues for each file. Finally, there are two optional parameters that the user may adjust. The parameter k is the desired number of clusters for grouping

the aligned fragments together, and σ is the noise parameter, which determines how strictly structural similarity is applied. For low σ-values, pFlexAna matches only fragments that are highly similar. For high σ-values, it also matches fragments with weak similarity and ignores small differences. Based on our experiences, σ-values between 0.2 and 0.4 work well for proteins with reasonably high-sequence homology (40% or higher). For proteins that are even more distantly related, we find σ-values up to 0.8 to be useful as well, though this may lead to false positive matches in some cases.

Output

pFlexAna identifies pairs of fragments—one fragment from each protein—that show significant conformational similarity. They are listed in a table and grouped into clusters, according to the similarity of the geometric transformation that aligns the fragment pair. See Figure 3 for an example. Two or more fragment pairs may move together and form a semi-rigid domain. By categorizing fragment pairs that are moving together versus moving independently, clustering helps to characterize conformational changes caused by domain movements or correlated motions.

For each aligned fragment pair, pFlexAna provides a static cartoon of the proteins superimposed according to the aligned pair. This highlights the movement of the proteins with respect to the aligned regions. Jmol links are also provided so that the user may view the proteins from different angles. Finally, the transformed PDB coordinate files that align the second protein to the first are available for download and the transformation matrices used for alignment are provided. The same information is also provided for each cluster of aligned fragment pairs.

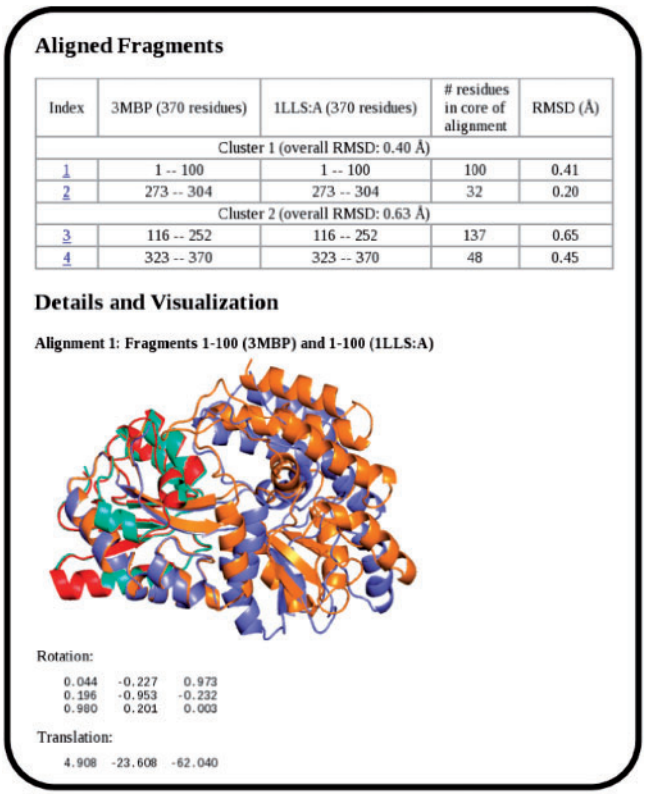

Figure 3. A screenshot showing the pFlexAna result for the maltose-binding protein. The pose of the protein is different from that in Figure 2.

Processing

To analyze the conformational changes between two protein structures, pFlexAna proceeds in two main steps. In the first step, it identifies pairs of fragments that are structurally similar. These fragment pairs serve as rigid cores for superimposing the two proteins and detecting conformational changes with respect to the cores. In the second step, pFlexAna clusters the aligned fragment pairs into 'super-alignments', according to the similarity of geometric transformations that align two fragments in a pair. The result is that the dominant conformational changes occur between the clusters, while the smaller conformational changes occur within a cluster. We now describe the two steps in greater detail below.

To identify structurally similar fragment pairs, pFlexAna uses an all-against-all approach that effectively compares every contiguous fragment from one protein with that from the other. First, we match the residues of the two proteins by 'sliding' one protein against the other (Figure 4). For each sliding position, let P and P' represent the matched residues from the two proteins, respectively. We generate all contiguous fragments of P and P' and check their structural similarity by applying a reliable statistical test (4). Briefly, this test uses a Gaussian noise model to represent acceptable deviations in atomic coordinates between two putatively similar protein fragments. Two fragments are considered statistically different if their coordinates deviate too much according to the Gaussian model. We then assign to each residue a flexibility score that incorporates information from all fragment comparisons that involve this residue. A high score indicates greater rigidity. Finally, each pair of aligned fragments must pass the following filter: the fragments must consist entirely of residues whose flexibility scores are above a threshold, and they must be longer than a given minimum length. Intuitively, this filter ensures that such a pair of fragments and all subfragments contained in them are structurally identical up to statistical variations. The minimum length requirement is imposed to avoid accidental structural matches. One advantage of our method is that it highlights the genuine conformational changes by suppressing the spurious ones due to noise. More details on this method and its advantages are available in ref. (4). Since we perform an

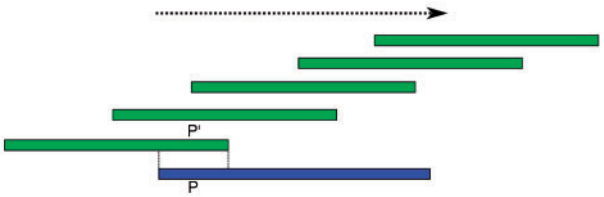

Figure 4. Sliding one protein against the other.

exhaustive comparison of all contiguous fragments in the two protein structures, our method can detect structurally similar fragments that are in opposite order along the protein sequences.

After obtaining an exhaustive list of structurally similar fragment pairs, we need to resolve the conflicts among them. A conflict occurs if in the list of matched fragment pairs, a residue from one protein is matched with multiple residues from the other protein. In practice, we have found that biologically significant matches have much longer fragment length than the spurious ones due to accidental structural similarity. So we remove the conflicts in the list by preferring fragment pairs with longer length.

Finally, pFlexAna hierarchically clusters the fragment pairs. The idea is to treat each structurally similar fragment pair as a point and divide these points into k clusters so that the dominant structural differences occur across clusters. To do this, we build a similarity graph. The vertices of this graph correspond to the fragment pairs obtained after conflict resolution. There is an edge between every two vertices, and the associated edge weight represents the similarity between the fragment pairs corresponding to the two vertices. Here, similarity is defined as the root mean square deviation (RMSD) for the best superposition when the two fragment pairs are combined. After constructing the similarity graph, we recursively remove from the graph the edge with the greatest weight and thus separate the most dissimilar fragment pairs, which represent the largest structural difference. We continue this edge removal process until the graph breaks into k connected components.

The web server implementation of our method uses PHP for its front-end interface, C++ for back-end processing, and a Ruby daemon to interface the front and back ends. The output images are generated using PyMol (15) and interactive displays of each alignment are provided using Jmol (16).

RESULTS

We illustrate the results of pFlexAna on several representative cases.

Ubiquitin ligase

The crystal structures of the E6AP and WWP1 ubiquitin ligases [PDB codes 1D5F:C (2) and 1ND7 (3)] display a striking example of domain movement (Figure 1). pFlexAna finds six core regions of significant structural similarity between the proteins and successfully clusters them into two domains which move independently (see Figure 1). These images are obtained directly from the web server, but annotated with an arrow and dotted ovals to draw attention to the domains that are aligned. The relative movement between the domains has been characterized as a 100 degree sweep with a 30 degree tilt (3). Superpositions, such as those shown in Figure 1, help to highlight the context of each protein region undergoing conformation change.

Superpositions of similar regions, as shown in Figure 1, are ideal for visualization because they show the difference as a divergence from a fixed reference point. Beyond simply showing the context of each 'flexible' region, each superposition indicates a transition from similar (rigid) to different (flexible).

Glutamine amidotransferase and arabinose-binding protein

The ubiquitin ligases have moderate sequence identity (<40%), but we have tested pFlexAna on a range of protein pairs with low sequence identity such as the isomerase domain of glucosamine 6P synthase [1MOQ (17)] and arabinose binding protein [8ABP (18)], which have 10% sequence identity. Figure 5 shows the alignments of 1MOQ and 8ABP. pFlexAna finds two independent domains. Each aligned domain centers around a parallel β-sheet buried by α-helices. In this figure, each alignment is rotated for viewing down the plane of an aligned sheet. To see the magnitude of the domain movement, note that the N-terminal domain of 8ABP (magenta) appears in the upper-right of both views, and 1MOQ (orange) is dramatically shifted between the two alignments.

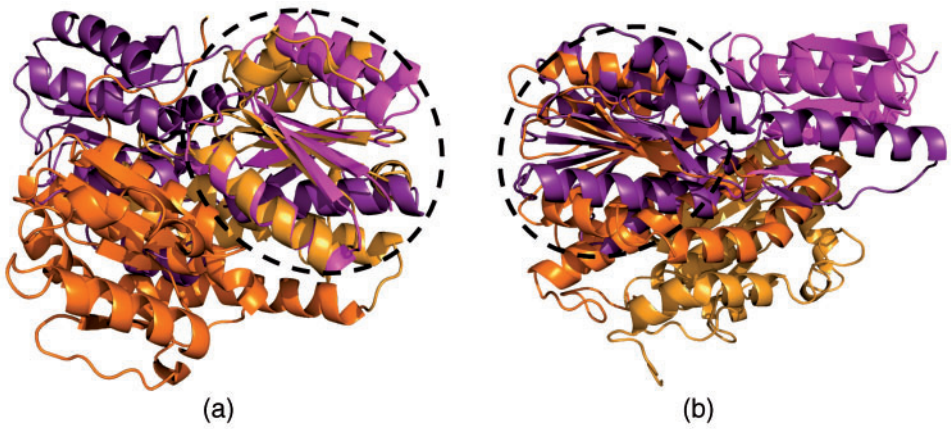

Figure 5. Proteins 1MOQ (orange) and 8ABP (purple) are aligned by pFlexAna according to their N-terminal domains in (**a**) and according to their C-terminal domains in (**b**). The N-terminal residues of each protein are shaded lighter than the C-terminal residues. Each alignment includes a parallel β-sheet (circled).

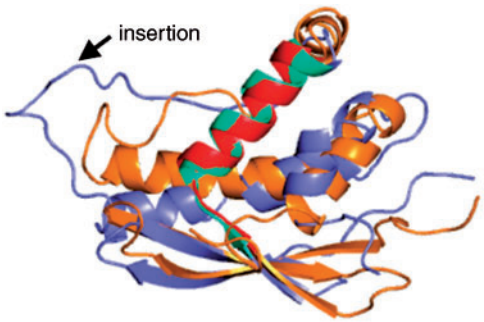

Figure 6. PX domain proteins have strong similarities at the N- and C-termini, but not in the loop region indicated by the arrow.

PX domain

pFlexAna is particularly useful for proteins exhibiting domain moments, but applying it to proteins with more subtle conformational changes and without major domain movements can also be informative. In particular, we set parameter $k = 1$ when we expect only a single cluster corresponding to the aligned protein core. For example, Figure 6 compares two PX domain proteins [1KMD (19) and 1GD5 (20)]. The N- and C-termini of the two proteins are conserved, but the loop region in between contains several insertions/deletions. Most structural alignment methods attempt to align these two proteins by superimposing their entire backbones globally. While this may be useful for some purposes, single global alignments may be skewed as they include the loop region. A more informative view of the evolutionary impact on conformational structure is shown by aligning only the conserved segments. On the PX domains, pFlexAna automatically finds the fragments of tightest similarity, distinct from the loop region, and it presents the clustered alignment result without skew from the inserted amino acids.

DISCUSSION

pFlexAna is a new tool for protein structure comparison. It detects and displays conformational changes in remotely related proteins, without relying on sequence homology. Our tests show that pFlexAna helps in analyzing a wide variety of conformational changes. It can detect conformational changes in proteins with identical sequences or those with very low sequence similarity. It can detect dramatic domain movements as well as smaller, more subtle conformational changes.

The output of pFlexAna helps the analysis of conformational changes in several ways. pFlexAna uses a reliable statistical test of structural similarity to demarcate the protein regions that undergo conformational changes. It also clusters the protein fragments that move in tandem. The clustering helps in finding domains comprised of fragments that are discontiguous and out of order along the sequences, e.g. in the case of the maltose-binding protein (Figure 3). Visualization of such information, provided in pFlexAna's output, helps the user to scan for hinge locations, active binding sites, domain movements, etc. For more detailed analysis, the user may use the quantitative information provided in pFlexAna's output, e.g. use the provided transformation matrix for alignment to compute the angle of bending at a particular hinge. pFlexAna's output may also help users interested in predicting conformational motion. They may use the information that pFlexAna provides as a starting point and combine it with other methods such as targeted molecular dynamics (21), where an MD simulation trajectory is 'directed' from an inital conformation towards an alternative target conformation.

pFlexAna uses purely structural information to match two proteins. While this is often advantageous as it does not rely on sequence homology, it can potentially cause false positive matching. For example, for proteins with many secondary structure elements (SSEs), it is possible for unrelated long alpha helices or other common motifs comprised of multiple SSEs to be identified as aligned fragment pairs. Comparing glutamine amidotransferase (1MOQ) with a ubiquitin ligase (1ND7) matches two pairs of highly similar α-helices, with RMSD values of 0.22 Å and 1.64 Å, respectively. While such aligned pairs are justified based on purely structural similarity, they may not be biologically significant to form a basis for inferring conformational changes. In the future, we intend to filter out such false positives by incorporating secondary structure or sequence information during the matching.

ACKNOWLEDGEMENTS

I.M. and C.V. gratefully acknowledge support by Biomedical Research Council of A*STAR, Singapore. L.T.K. was supported by a Lee Kuan Yew Postdoctoral Fellowship. Funding to pay the Open Access publication charges for this article was provided by the NUS AcRF grant R-252-000-342-112.

Conflict of interest statement. None declared.

REFERENCES

1. Swain,J.F. and Gierasch,L.M. (2006) The changing landscape of protein allostery. *Curr. Opin. Struct. Biol.*, **16**, 102–108.
2. Huang,L., Kinnucan,E., Wang,G., Beaudenon,S., Howley,P.M., Huibregtse,J.M. and Pavletich,N.P. (1999) Structure of an e6ap-ubch7 complex: insights into ubiquitination by the e2-e3 enzyme cascade. *Science*, **286**, 1321–1326.
3. Verdecia,M.A., Joazeiro,N., Wells,N.J., Ferrer,J.-L., Bowman,M.E., Hunter,T. and Noel,J.P. (2003) Conformational flexibility underlies ubiquitin ligation mediated by the wwp1 hect domain e3 ligase. *Mol. Cell*, **11**, 249–259.
4. Nigham,A. and Hsu,D. (2007) Protein conformational flexibility analysis with noisy data. In *Proc. ACM Int. Conf. on Computational Biology (RECOMB)*, Springer-Verlag, Berlin, pp. 396–411. http://motion.comp.nus.edu.sg/papers/recomb07.pdf.
5. Roccatano,D., Mark,A.E. and Hayward,S. (2001) Investigation of the mechanism of domain closure in citrate synthase by molecular dynamics simulation. *J. Mol. Biol.*, **310**, 1039–1053.
6. Gerstein,M. and Chothia,C. (1991) Analysis of protein loop closure. two types of hinges produce one motion in lactate dehydrogenase. *J. Mol. Biol.*, **220**, 133–149.
7. Wriggers,W. and Schulten,K. (1997) Protein domain movements: detection of rigid domains and visualization of hinges in comparisons of atomic coordinates. *Proteins Struct. Funct. Genet.*, **29**, 1–14.

8. Hayward,S. and Berendsen,H.J.C. (1998) Systematic analysis of domain motions in proteins from conformational change: new results on citrate synthase and t4 lysozyme. *Proteins Struct. Funct. Genet.*, **30**, 144–154.
9. Qi,G., Lee,R. and Hayward,S. (2005) A comprehensive and non-redundant database of protein domain movements. *Bioinformatics*, **21**, 2832–2838.
10. Quiocho,F.A., Spurlino,J.C. and Rodseth,L.E. (1997) Extensive features of tight oligosaccharide binding revealed in high-resolution structures of the maltodextrin transport/chemosensory receptor. *Structure*, **5**, 997–1015.
11. Rubin,S.M., Lee,S.Y., Ruiz,E.J., Pines,A. and Wemmer,D.E. (2002) Detection and characterization of xenon-binding sites in proteins by 129xe nmr spectroscopy. *J. Mol. Biol.*, **322**, 425–440.
12. Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (ce) of the optimal path. *Protein Eng.*, **11**, 739–747.
13. Ye,Y. and Godzik,A. (2003) Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics*, **19** (Suppl 2), ii246–ii255.
14. Shatsky,M., Wolfson,H.J. and Nussinov,R. (2002) Flexible protein alignment and hinge detection. *Proteins Struct. Funct. Genet.*, **48**, 242–256.
15. Delano,W.L. (2002) *The PyMOL User's Manual*. DeLano Scientific, Palo Alto, CA.
16. Herraez,A. (2006) Biomolecules in the computer: Jmol to the rescue. *Biochem. Mol. Biol. Educat.*, **34**, 255–261.
17. Teplyakov,A., Obmolova,G., Badet-Denisot,M.A., Badet,B. and Polikarpov,I. (1998) Involvement of the c terminus in intramolecular nitrogen channeling in glucosamine 6-phosphate synthase: evidence from a 1.6 a crystal structure of the isomerase domain. *Structure*, **6**, 1047–1055.
18. Vermersch,P.S., Lemon,D.D., Tesmer,J.J. and Quiocho,F.A. (1991) Sugar-binding and crystallographic studies of an arabinose-binding protein mutant (met108leu) that exhibits enhanced affinity and altered specificity. *Biochemistry*, **30**, 6861–6866.
19. Lu,J., Garcia,J., Dulubova,I., Südhof,T.C. and Rizo,J. (2002) Solution structure of the vam7p px domain. *Biochemistry*, **41**, 5956–5962.
20. Hiroaki,H., Ago,T., Ito,T., Sumimoto,H. and Kohda,D. (2001) Solution structure of the px domain, a target of the sh3 domain. *Nat. Struct. Biol.*, **8**, 526–530.
21. van der Vaart,A. and Karplus,M. (2005) Simulation of conformational transitions by the restricted perturbation-targeted molecular dynamics method. *J. Chem. Phys.*, **122**, 114903.