Advanced pairwise structure alignments of proteins and analysis of conformational changes

Maria Elena Ochagavia\textsuperscript{1,2}, Jean Richelle\textsuperscript{2} and Shoshana J. Wodak\textsuperscript{2,*}

\textsuperscript{1}Center for Genetic Engineering and Biotechnology, Ave. 31 between Avenues 158 and 190, Cubanacán, City of Havana, 10600 Cuba and \textsuperscript{2}Service de Conformation de Macromolécules Bioligiques et de Bioinformatique CP 160/16, P2, Université Libre de Bruxelles, Ave. F. Roosevelt, 1050 Bruxelles, Belgium

Received on July 18, 2001; revised on October 30, 2001; accepted on November 14, 2001

ABSTRACT
Motivation: Comparing the 3D structures of two proteins or analyzing the structural changes undergone by a protein upon ligand binding or when it crystallizes under different conditions, can be both tricky and tedious, especially when the two proteins are distantly related, or when the structural changes are complex. Readily accessible tools for performing these tasks automatically and reliably should therefore be welcome.

Results: We describe a web interface to several automatic procedures for performing pairwise structure superposition in a flexible manner, for detailed analyses of conformational changes and for displaying the results in a pictorial fashion.

Availability: This interface can be accessed at the Brus-

Contact: shosh@ucmb.ulb.ac.be

INTRODUCTION
In recent years a battery of automatic procedures has been developed for optimally superimposing the atomic coordinates of pairs of protein structures in the absence of prior knowledge on the equivalent residues, also termed structural alignment (for review see Lemmen and Lengauer, 2000). Several methods are also available for efficiently searching structural databases for proteins with similar folds to a given protein structure (Holm and Sander, 1994). Such methods have been combined with clustering techniques to generate protein fold classifications (CATH; Orengo et al., 1997, SCOP; Murzin et al., 1995) or used to provide information on structural neighbors (FSSP; Holm and Sander, 1996, VAST; Gibrat et al., 1996, Combinatorial Extension (CE); Shindyalov and Bourne, 1998).

Several services for performing pairwise structural alignments of proteins and for finding structural neighbors can be found on the world wide web (a list see http://www.ucmb.ulb.ac.be/SCMBB/Tools.html). But most of these offer only limited access to the programs by imposing default settings that cannot be altered. This often poses problems for the alignment of structurally more divergent proteins, where small adjustments of the parameters may be necessary to find meaningful solutions.

An important application of structure comparison methods, which has received comparatively little attention, is the analysis of protein conformational changes, whose object is to provide a detailed description of the structural deformations that lead from one protein conformation to another (Chothia et al., 1983; Lesk and Chothia, 1984). Such analyses, typically applied to structures of proteins determined under different experimental conditions, or in the presence and absence of a bound ligand, are used to infer protein motion associated with function, or to analyze conformational flexibility. Examples abound on the valuable insight provided by these studies on the movements of protein loops, domains and whole subunits, associated with catalysis, allosteric regulation, or DNA, RNA and protein recognition. But most of these relied on a semi-manual approach involving repeated applications of simple rigid-body coordinate superposition (Kabsch, 1978; McLachlan, 1979) to different subsets of atoms, in order to identify relative movements of sub-structures.

With the drastic increase in the pace of protein structure determination, expected to result from the various structural genomics initiatives (Brenner and Levitt, 2000), it becomes paramount to use more efficient computational tools also for the type of analyses described above. A few computational procedures for analyzing and characterizing protein motion have been available for some time. But
they have not been widely used, due in part to the lack of easy access.

The sieve–fit program (Lesk, 1991), identifies groups of residues that behave as a rigid body. But to detect several rigid-bodies that move relative to one another, not an uncommon observation, the procedure must be run repeatedly, using different subsets of atoms specified manually by the user. The fit-all algorithm (Gerstein and Chothia, 1991), gives indications on whether a conformational change involves a hinge or shear movement. It performs a series of superpositions, and generates a 2D plot from which the hinge residues, when they exist, can be deduced visually. These and other tools have been used to set up the database of protein motions (Gerstein and Krebs, 1998; Krebs and Gerstein, 2000).

Here we present a new web interface to automatic procedures for optimally aligning two protein structures in a flexible manner and for performing a systematic analysis of the differences between their backbone structures. Both procedures have been extensively used in our laboratory for various applications, such as the identification of structural neighbors in the CASP1 assessment of the threading predictions (Lemer et al., 1995), and to describe the conformational changes in a variety of protein systems.

AUTOMATIC ANALYSIS OF CONFORMATIONAL CHANGES

The procedure for analyzing conformational changes between two proteins (Boutonnet et al., 1995a,b), is an extension of the structure alignment algorithm described above. In a first step it computes the optimal global alignment of the two protein structures, by systematically superimposing backbone segments in secondary structure elements and loops and assembling them by order of decreasing structural similarity. The assembly is performed using a multiple linkage clustering procedure in which each group of segments can be combined more than once with any other group, yielding a set of intertwined clustering trees, which are stored. In a second step, these trees are analyzed completely automatically, in order to describe several features of the conformational change. One is the static core, defined as the parts of the two structures, which remain unchanged (Lesk and Chothia, 1984; Gerstein and Chothia, 1991). Another is the movers, the secondary structure and loops segments that move relative to the static core (Boutonnet et al., 1995a,b). Options are also available to perform a systematic analysis of substructures that undergo limited deformations, while moving relatively to each other, the rigid bodies.

THE MECHANISM OF DOMAIN MOTION IN GLNP

To illustrate the analysis of the conformational changes performed with our server, we apply it to Glutamine Binding Protein (GlnBP; Nohno et al., 1986). The 3D structures of both the ligand-free ‘open conformation’ (PDB ID: 1ggg; Hsiao et al., 1996) and the complex with glutamine ‘closed-cleft conformation’ (PDB ID: 1wdr; Sun et al., 1998) have been determined by x-ray diffraction. GlnBP is composed of two domains, linked by two β-strands, hydrogen bonded to each other. The structural change between the two forms can be described as a large-scale movement of the two domains, mediated by a hinge involving large local backbone deformations of the two connecting strands (Sun et al., 1998).

Using our server, the conformational change analysis was applied to residues 6–221 in chain A of both proteins and the Nuclei Growth Path (NGP) (Figure 1) was computed. This is a particular path followed along the clustering trees describing the fragment assembly hierarchy used in constructing the global alignment. It is obtained by moving along the nodes of the trees, from the leaves to the root and joining two nodes (clusters) whenever the rmsd value of the joined nodes is less than a specified threshold. The protein segments joined at each step along the NGP are displayed as horizontal bars in the plot of Figure 1 (red bars are α-helices and green bars, β-strands). The horizontal axis in this
Fig. 1. Graphic output of the NGP analysis of the conformational changes between the open and closed forms of the GlnBP. In this plot, the protein segments joined at each step along the NGP are displayed as horizontal bars. Segments in α-helices (H1–H8) appear as red bars, β-strands (S1–S12), as green bars. These segments involve the following residues from chain A, respectively: S1 (7–9), S2 (17–20), S3 (23–25), H1 (27–38), S4 (45–47), H2 (53–58), S5 (64–65), H3 (73–76), S6 (80–81), S7 (86–95), S8 (111–115), H4 (119–127), S9 (132–136), H5 (139–147), S10 (153–157), H6 (158–166), S11 (173–185), H7 (194–209), H8 (212–220). Given in parentheses are PDB residue numbers. The horizontal axis indicates the positions of the assembled segments along the polypeptide. The vertical axis shows the rmsd computed when all the segments with the same vertical coordinate are simultaneously superimposed (left) and lists the step number along the NGP (right). The conformational changes analysis was performed with default parameters on residues 6–221 in chain A of the open and closed conformations of GlnBP, PDB codes 1ggg and 1wdn, respectively.

The NGP analysis of GlnBP shows that the first 11 nodes (steps) visited along the clustering trees assemble 12 segments (totaling 66 residues) from the N- and C-terminal portions of the polypeptide, and that these segments superimpose with a relatively low rmsd of 0.52 Å. Addition of the 8 segments from the middle portion of polypeptide (step 12) is accompanied by a large rmsd jump (4.1 Å). This is a clear indication that the structure has two domains, which move relative to each other. The large rmsd jump is indicative of a hinge mechanism; a shear mechanism would be characterized by a series of more closely spaced lines along the vertical axis, each grouping a smaller set of segments. The identified domains correspond exactly to those described by the crystallographers (Sun et al., 1998). Moreover, in the NGP plot the first and the last β-strands added in step 12, are the two hinge strands connecting the two domains (represented as ribbons in Figure 2), as previously reported (Sun et al., 1998).

While investigating the domain motion in GlnBP, it may also be useful to identify the part of the structure that remains unchanged, the static core. To identify such core, the Jump Minimizing Path (JMP) along the clustering trees must be examined. This path starts at the leaves of the tree, and then travels down the branches, while minimizing the difference in rmsd between any two nodes. The static core is defined as the node that is separated by a specified difference in rmsd from the next node along this path.

In GlnBP, the static core is formed by 11 segments. Seven in β-strands S1–S6, and S12, and four in α-helices H1–H3 and H8, all belonging to the large structural domain. They total 50 residues, which superimpose with
a rmsd of 0.41 Å. Superimposing the static core residues in the two GlnBP structures illustrates well the domain motion (Figure 2).

The rigid-body analysis reveals two main rigid bodies. One is in the small domain and comprises a total of 19 residues in strands S8, S10, and helix H6, which superimpose with a rmsd of 0.27 Å. The other is in the large domain. It comprises all the secondary structure elements of the static core, except for helix H3, but contains in addition helix H7, which is only slightly displaced relative to the static core, leading to a somewhat higher rmsd value for this rigid substructure (0.49 Å), than that of the static core, given above.

Finally, the program also lists the individual pairs of aligned segments (data not shown). This provides information on the structural deformation undergone locally by individual segments along the chain. Interestingly, the structurally most similar segments in the small domain are S8 (rmsd = 0.08) and S10 (rmsd = 0.15), which contain several ligand binding residues (Lys 115, His 156, Asp 157). The most similar segments in the large domain are in strands S4–S6 and S12, located at the heart of the static core.

ACKNOWLEDGEMENTS

M.E.O. is indebted to the Fondation Brachet, for travel support. This work was supported by the Action de Recherches concertées de la Communauté Française de Belgique, project no 97/01-211 and by the Center for Genetic Engineering and Biotechnology, Havana Cuba.

REFERENCES