W590–W594 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh477

SuperPose: a simple server for sophisticated structural superposition

Rajarshi Maiti, Gary H. Van Domselaar, Haiyan Zhang and David S. Wishart*

Departments of Biological Sciences and Computing Science, University of Alberta, Edmonton, AB, T6G 2E8, Canada

Received February 15, 2004; Revised April 14, 2004; Accepted May 3, 2004

ABSTRACT

The SuperPose web server rapidly and robustly calculates both pairwise and multiple protein structure superpositions using a modified quaternion eigenvalue approach. SuperPose generates sequence alignments, structure alignments, PDB (Protein Data Bank) coordinates and RMSD statistics, as well as difference distance plots and images (both static and interactive) of the superimposed molecules. SuperPose employs a simple interface that requires only PDB files or accession numbers as input. All other superposition decisions are made by the program. SuperPose is uniquely able to superimpose structures that differ substantially in sequence, size or shape. It is also capable of handling a much larger range of superposition queries and situations than many standalone programs and yields results that are intuitively more in agreement with known biological or structural data. The SuperPose web server is freely accessible at http://wishart.biology.ualberta.ca/SuperPose/.

INTRODUCTION

In the same way that sequence comparisons can provide tremendous insight into the origins, function, location, interactions and activity of a given protein, so too can structure comparison. In fact, because structure is actually much more conserved than sequence, structure comparisons allow us to look even further back into biological prehistory (1). The most common method for three-dimensional (3D) structure comparison is called structure superposition. Superposition or superimposition is simply the process of rotating or orienting an object until it can be directly overlaid on top of a similar object. While quite simple in principle, in practice superposition is actually quite difficult. Most modern superposition methods employ complex computational methods such as Lagrangian multipliers, quaternion methods and matrix diagonalization techniques (2–4). Throughout the 1970s and 1980s, a number of standalone computer programs were described that employed these methods (2–5). More recently, structure superposition programs have found their way into commercial visualization packages such as those offered by Tripos and Accelrys. Additionally, several freeware modeling programs, including DeepView (6) and MolMol (7), are also capable of performing and visualizing certain kinds of molecular superpositions. However, in order to perform most molecular superpositions with either commercial or freeware products, users must become quite familiar with some rather complex interfaces. Furthermore, not all of these standalone packages are compatible with common operating systems or compilers used in many molecular biology or teaching labs.

As an active structural biology laboratory, we have experienced a good deal of frustration in trying to use both commercial and freeware products to superimpose two or more molecules in a consistent and reliable way. Beyond the continuing problems of overly complex graphical user interface (GUI) design and operating system incompatibility we have frequently found that the reported RMSD (root mean square deviation) values differ substantially between packages or that the RMSD values are incompletely described. Furthermore, some packages are quite restrictive in what can be superimposed (structures must be of identical length), how many molecules can be superimposed (most permit just two molecules to be superimposed), how the molecules are superimposed and how different two structures can be when superimposed.

To overcome these ongoing problems we have gone back to the drawing board and written a robust macromolecular superposition web server—called SuperPose—that appears to overcome the underlying problems among standalone programs regarding GUI complexity, platform incompatibility, RMSD inconsistency, sequence length compatibility and limited superposition capability. SuperPose takes PDB (Protein Data Bank) files or PDB accession numbers as input and rapidly generates textual, numeric and visual output that fully describes two or more superimposed molecules. SuperPose was specifically designed to permit macromolecular superposition to be easily done and displayed by anyone, anywhere at any time.

*To whom correspondence should be addressed. Tel: +1 780 492 0383; Fax: +1 780 492 1071; Email: david.wishart@ualberta.ca

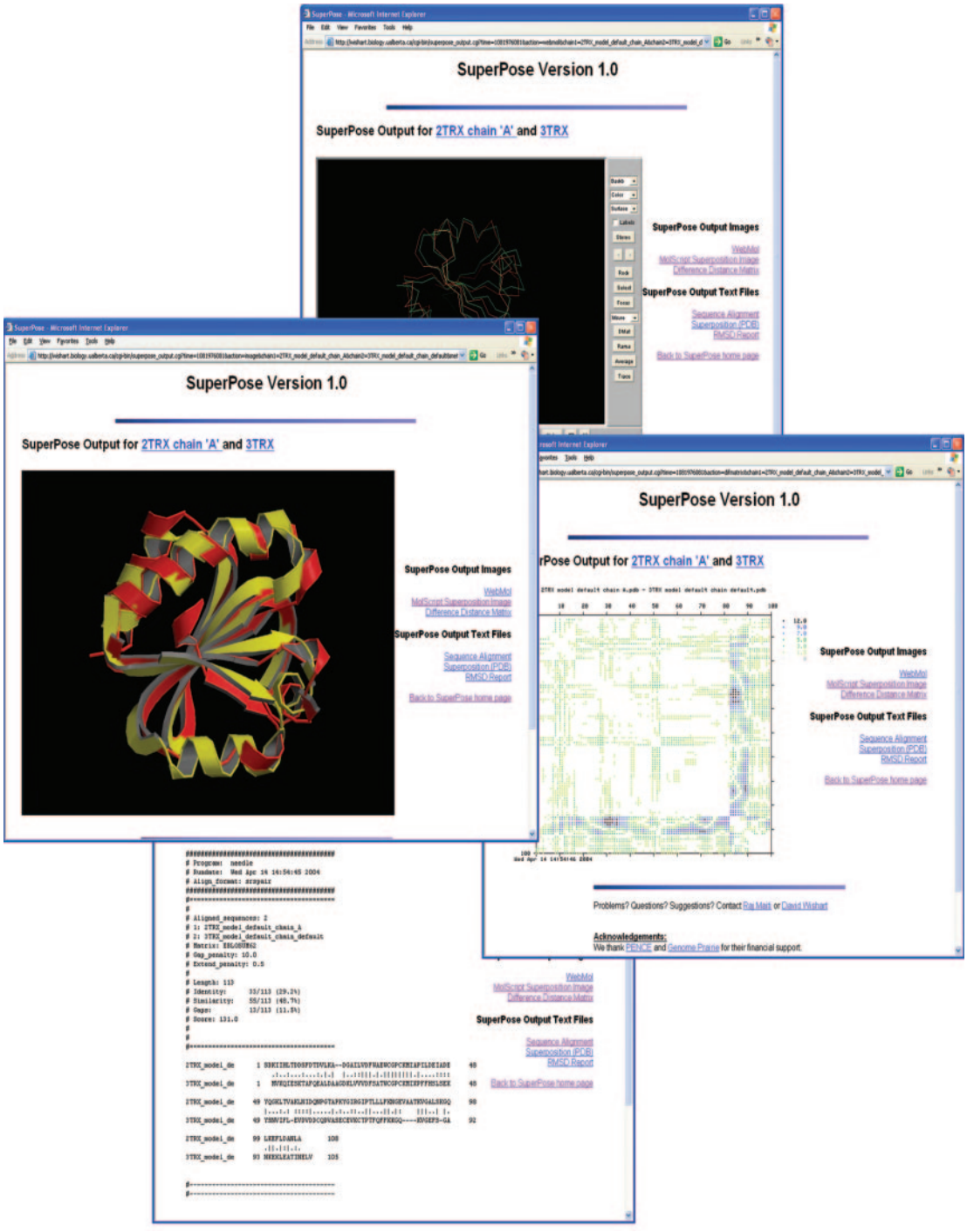

Figure 1. A screenshot montage of SuperPose server output showing the different kinds of graphical and textual output available. Shown are the WebMol viewer, a MolScript image, a pairwise alignment, a difference distance matrix and the RMSD output for a pairwise superposition of 2TRX_A and 3TRX_A (*Eschericia coli* thioredoxin and human thioredoxin). The two proteins have 29% sequence identity.

PROGRAM DESCRIPTION

SuperPose is composed of two parts, a front-end web interface (written in Perl and HTML) and a back-end for alignment, superposition, RMSD calculation and rendering (written in Perl and C). The front-end accepts two kinds of input, PDB text files (from a user's hard drive) or PDB accession numbers or any combination of both. If users choose to use PDB accession numbers, they can also designate which chain(s) they would like (2TRX_A means the A chain of 2TRX) in the input text box. Once the PDB accession numbers are chosen, the program automatically goes to the PDB website and retrieves the necessary files. SuperPose also allows users to interactively select chains within PDB files if they are not familiar with the chain structure or chain content of their

Table 1. Comparison of superposition statistics [backbone RMSD (Å) and residue matches] for SuperPose, DeepView and MolMol using superposition tasks of varying difficulty done without user intervention

Category Structure(s)—% sequence identity (ID)	SuperPose	DeepView	MolMol
Same sequence and similar structure (pair)			
Thioredoxin (2TRX_A on 2TRX_B)—100% ID	0.77 (1–108)	0.77 (1–108)	0.66 (1–108)
Hemoglobin (4HHB_A on 1DKE_A)—100% ID	0.37 (1–141)	0.37 (1–141)	—[a]
P21 Oncogene(6Q21_A on 6Q21_B)—100% ID	1.27 (1–171)	1.27 (1–171)	1.16 (1–171)
∼Same sequence and different structure (pair)			
Calmodulin (1A29 on 1CLL)—98.6% ID	0.82 (5–75) 23.83 (5–146)	15.02 (4–146)[b]	—[a]
Maltose Bind Prot. (1OMP on 1ANF)—100% ID	0.83 (1–112) 8.87 (2–370)	3.76 (1–370)[b]	3.76 (1–370)[b]
Similar structure and different length (pair)			
Hemoglobin (4HHB_A on 4HHB_B)—43% ID	1.61 (98.6% aligned)	1.21 (65% aligned)	—[a]
Thioredoxin (3TRX on 2TRX_A)—29% ID	4.23 (85.7% aligned)	1.70 (25.7% aligned)	—[a]
Lysozyme/Lactalbumin (1DPX on 1A4V)—36% ID	1.63 (99% aligned)	2.05 (55% aligned)	—[a]
Calmodulin/TnC (1CLL on 5TNC)—47% ID	6.83 (100% aligned)	5.24 (52% aligned)	—[a]
Similar structure and very different sequence			
Ubiquitin/Elongin (1UBI on 1VCB_A)—26% ID	3.22 (96% aligned)	2.19 (72.3% aligned)	—[a]
Thio/Glutaredoxin (3TRX on 3GRX_A)—7% ID	4.64 (92.7% aligned)	1.76 (14.6% aligned)	—[a]
Hemoglobins (1ASH on 2LHB)—17% ID	4.11 (93.8% aligned)	1.90 (19.7% aligned)	—[a]
Thioredoxins (1NHO_A on 1DE2_A)—22% ID	7.77 (77.6% aligned)	4.04 (21.2% aligned)	—[a]
Multiple structures and same sequence			
Pointed domain (1BQV, 28 chains)—100% ID	6.28 (100% aligned)	—[c]	5.88 (100% aligned)
Trypsin inhibitor (1PIT, 20 chains)—100% ID	1.32 (100% aligned)	—[c]	1.30 (100% aligned)
Oligomerization domain (1OLG, 4 chains)—100% ID	0.57 (100% aligned)	—[c]	0.58 (100% aligned)
Oxidoreductase (1NHO, 20 chains)—100% ID	0.96 (100% aligned)	—[c]	0.96 (100% aligned)

[a]Failed: atom mismatch.
[b]Could not detect hinge region.
[c]Failed.

chosen PDB file. This is done simply by clicking on the name of the chain in the scroll boxes that SuperPose generates after it has read each PDB file. To support alternative displays and alternative superpositions or to override SuperPose decisions, SuperPose offers three sets of options: (i) output options; (ii) alignment options and (iii) advanced options, which are listed below the SuperPose input form. Normally most users would have no need to change the default values. Nevertheless, detailed descriptions about what these options mean and how to fill out the option boxes are provided on both the SuperPose home page and its Help pages.

SuperPose is designed to handle five kinds of macromolecular superposition requirements: (i) superposition of two or more molecules of identical sequence but slightly different structure; (ii) superposition of two molecules of identical sequence but profoundly different structure (e.g. open and closed forms of calmodulin); (iii) superposition of two or more molecules of modestly dissimilar sequence, length and structure; (iv) superposition of two or more molecules with profoundly different lengths but similar structure or sequence; and (v) superposition of two or more molecules that are profoundly different in sequence but similar in structure. The most common scenario, and the one supported by most superposition packages, is scenario (i). This type of superposition is frequently done in generating NMR (nuclear magnetic resonance) structure ensembles, in comparing ligand-bound and ligand-free molecules and in comparing two different crystal isoforms. In scenario (i) sequence and sequence length differences are irrelevant and the problem can be framed as a pure geometrical optimization problem. However, for the other four scenarios, sequence and length information are relevant—as is information about local structure similarity. Unfortunately, most available superposition packages do not account for this kind of information and so they frequently perform poorly or require considerable user knowledge or input to get them to perform well. To deal with all five superposition scenarios, SuperPose employs a combination of four techniques: (i) pairwise or multiple pairwise sequence alignment; (ii) secondary structure alignment (when sequence identity <25%); (iii) difference distance matrix calculation; and (iv) quaternion superposition.

Beginning with an input PDB file or set of files, SuperPose first extracts the sequences of all chains in the file(s). Each sequence pair is then aligned using a Needleman–Wunsch pairwise alignment algorithm (8) employing a BLOSUM62 scoring matrix. If the pairwise sequence identity falls below the default threshold (25%), SuperPose determines the secondary structure using VADAR (volume, area, dihedral angle reproter) (9) and performs a secondary structure alignment using a modified Needleman–Wunsch algorithm. After the sequence or secondary structure alignment is complete, SuperPose then generates a difference distance (DD) matrix (10) between aligned alpha carbon atoms. A difference distance matrix can be generated by first calculating the distances between all pairs of Cα atoms in one molecule to generate an initial distance matrix. A second pairwise distance matrix is generated for the second molecule and, for equivalent/aligned Cα atoms, the two matrices are subtracted from one another, yielding the DD matrix. From the DD matrix it is possible to quantitatively assess the structural similarity/dissimilarity between two structures. In fact, the difference distance method is

Table 2. Comparison of superposition RMSD values [backbone RMSD (Å) and match residues] for SuperPose, DeepView and MolMol using a 'forced' superposition of matching residues identified by DeepView's Magic Fit

Category Structure(s)—% Sequence identity (ID)	SuperPose	DeepView	MolMol
Same sequence and similar structure (pair)			
Thioredoxin (2TRX_A versus 2TRX_B)—100% ID	0.77 (1–108)	0.77 (1–108)	0.66 (1–108)
Hemoglobin (4HHB_A versus 1DKE_A)—100% ID	0.37 (1–141)	0.37 (1–141)	0.36 (1–141)
P21 Oncogene(6Q21_A versus 6Q21_B)—100% ID	1.27 (1–171)	1.27 (1–171)	1.16 (1–171)
∼Same sequence and different structure (pair)			
Calmodulin (1A29 versus 1CLL)—98.6% ID	14.97 (4–146)	15.02 (4–146)	14.94 (4–146)
Maltose Bind Prot. (1OMP versus 1ANF)—100% ID	3.76 (1–370)	3.76 (1–370)	3.76 (1–370)
Similar structure and different length (pair)			
Hemoglobin (4HHB_A versus 4HHB_B)—43% ID	1.21 4HHB_A (51–141) 4HHB_B (56–146)	1.21 4HHB_A (51–141) 4HHB_B (56–146)	1.16 4HHB_A (51–141) 4HHB_B (56–146)
Thioredoxin (3TRX versus 2TRX_A)—29% ID	1.70 3TRX (23–49) 2TRX_A (23–49)	1.70 3TRX (23–49) 2TRX_A (23–49)	1.63 3TRX (23–49) 2TRX_A (23–49)
Lysozyme/Lactalbumin(1DPX versus 1A4V)—36% ID	2.05 1DPX (32–99) 1A4V (29–96)	2.05 1DPX (32–99) 1A4V (29–96)	1.91 1DPX (32–99) 1A4V (29–96)
Calmodulin/TnC (1CLL versus 5TNC)—47% ID	5.24 1CLL (4–78) 5TNC (14–88)	5.24 1CLL (4–78) 5TNC (14–88)	5.21 1CLL (4–78) 5TNC (14–88)
Similar structure and very different sequence			
Ubiquitin/Elongin (1UBI versus 1VCB_A)—26% ID	2.19 1UBI (12–66) 1VCB_A (13–67)	2.19 1UBI (12–66) 1VCB_A (13–67)	2.11 1UBI (12–66) 1VCB_A (13–67)
Thio/Glutaredoxin(3TRX versus 3GRX_A)—7% ID	1.75 3TRX (78–89) 3GRX_A (54–65)	1.76 3TRX (78–89) 3GRX_A (54–65)	1.51 3TRX (78–89) 3GRX_A (54–65)
Hemoglobins (1ASH versus 2LHB)—17% ID	1.90 1ASH (26–54) 2LHB (34–62)	1.90 1ASH (26–54) 2LHB (34–62)	1.76 1ASH (26–54) 2LHB (34–62)
Thioredoxins (1NHO_A versus 1DE2_A)—22% ID	4.04 1NHO_A (13–30) 1DE2_A (14–31)	4.04 1NHO_A (13–30) 1DE2_A (14–31)	3.85 1NHO_A (13–30) 1DE2_A (14–31)
Multiple structures and same sequence			
Pointed domain (1BQV, 28 chains)—100% ID	6.28 (100% aligned)	Failed	5.88 (100% aligned)
Trypsin inhibitor (1PIT, 20 chains)—100% ID	1.32 (100% aligned)	Failed	1.30 (100% aligned)
Oligomerization domain (1OLG, 4 chains)—100% ID	0.57 (100% aligned)	Failed	0.58 (100% aligned)
Oxidoreductase (1NHO, 20 chains)—100% ID	0.96 (100% aligned)	Failed	0.96 (100% aligned)

particularly good at detecting domain or hinge motions in proteins [see scenario (ii)]. SuperPose analyzes the DD matrices and identifies the largest contiguous domain between the two molecules that exhibits <2.0 Å difference. From the information derived from the sequence alignment and DD comparison, the program then makes a decision regarding which regions should be superimposed and which atoms should be counted in calculating the RMSD. This information is then fed into the quaternion superposition algorithm and the RMSD calculation subroutine. The quaternion superposition program is written in C and is based on both Kearsley's method (4) and the PDBSUP Fortran program developed by Rupp and Parkin (11). Quaternions were developed by W. Hamilton (the mathematician/physicist) in 1843 as a convenient way to parameterize rotations in a simple algebraic fashion. Because algebraic expressions are more rapidly calculable than trigonometric expressions using computers, the quaternion approach is exceedingly fast.

SuperPose can calculate both pairwise and multiple structure superpositions [using standard hierarchical methods (5)] and can generate a variety of RMSD values for alpha carbons, backbone atoms, heavy atoms and all atoms (average and pairwise). When identical sequences are compared, SuperPose also generates 'per residue' RMSD tables and plots to allow users to identify, assess and view individual residue displacements.

OUTPUT AND DISCUSSION

SuperPose produces up to seven kinds of output: (i) a PDB file containing the coordinates of the superimposed molecules; (ii) a PDB file containing the backbone coordinates of a single averaged structure (if all sequences match identically); (iii) a sequence or secondary structure alignment (pairwise or multiple) file of the sequences used in the alignment; (iv) a difference distance matrix (if only two molecules are superimposed); (v) an RMSD report that contains the calculated RMSD values (in Angstroms) between the superimposed molecules; (vi) a still image (PNG, portable network graphics file) of the superimposed molecules generated using MolScript (12); and (vii) a WebMol (13) applet view of the superimposed molecules. A montage illustrating the kinds of output produced by SuperPose is shown in Figure 1. By default, SuperPose produces the WebMol image of the superimposed

structures and a set of hyperlinks located on the right side of the screen. All other data (RMSD values, PDB files, images, and so on) can be accessed, saved or viewed by following the hyperlinks on the SuperPose output page. Images and textual output generated from SuperPose can also be saved or copied directly to the user's hard disk or loaded into standard presentation or molecular visualization programs. The data generated by SuperPose are kept in a temporary folder and are not stored for more than 24 h on the server. Similarly, SuperPose sessions that are inactive for >20 min are terminated. As far as we are aware, there is only one other operational structure superposition server, ProSup (14). ProSup is designed only to identify and superimpose small regions of proteins that are structurally very similar (<1.1 Å RMSD). Further, ProSup performs only pairwise structural comparisons, not multiple structure superpositions.

SuperPose appears to be unique as both a general superposition server and in its ability to handle difficult superposition tasks. Additionally, SuperPose provides a wide range of interactive viewing options (color, black-and-white, stereo, mono, ribbon, backbone), file (text/image) outputs and RMSD outputs not found on other servers or in other standalone packages. As seen in Table 1, tests performed on a number of 'challenging' superposition tasks (1A29 on 1CLL; 5TNC on 1CLL; 2TRX on 3TRX; ensemble superpositions of 28 ETS pointed domains—1BQV) indicate that SuperPose is able to automatically identify hinge motions, perform superpositions with structures of very different lengths or atom numbers, correctly superimpose very remotely related structures and consistently align structures over a much larger portion of their sequence length than either DeepView or MolMol. Interestingly, many of the superposition tasks listed in Table 2 were either not possible (MolMol requires exact matches of atom/residue numbers) or yielded non-sensical results (DeepView aligns to the longest contiguous matching segment—regardless of length). However, when the same residues are matched and assessed (a 'forced' or manual superposition), SuperPose is able to reproduce RMSDs for both pairwise and multiple structure superpositions that agree well with those values reported by DeepView and MolMol (Table 2). The one difference appears to lie in the fact that MolMol ignores carbonyl oxygen atoms in its evaluation of backbone RMSD values.

In summary, SuperPose provides a simple-to-use, web-accessible approach to performing a wide range of sophisticated structural superpositions. It is unique in that it combines sequence alignment and difference distance matrix calculations to constrain its quaternion eigenvalue superposition calculations. SuperPose has been designed to provide an abundance of useful textual and visual outputs that allow both structural 'novices' and experienced structural biologists to explore and compare complex protein structures. Our hope is that SuperPose will make structural superposition far more accessible and far simpler than it currently is. We also hope that it will eventually find a role in teaching life science students about the importance of structural comparisons. The SuperPose web server is accessible at http://wishart.biology.ualberta.ca/SuperPose/.

ACKNOWLEDGEMENTS

Funding for this project was provided by the Protein Engineering Network of Centres of Excellence (PENCE Inc.) and Genome Prairie (a division of Genome Canada).

REFERENCES

1. Orengo,C.A., Pearl,F.M. and Thornton,J.M. (2003) The CATH domain structure database. *Methods Biochem. Anal.*, **44**, 249–271.
2. MacLachlan,A.D. (1982) Rapid comparison of protein structures. *Acta Crystallogr. A.*, **38**, 871–873.
3. Kabsch,W.A. (1978) Discussion of solution for best rotation of two vectors. *Acta Crystallogr. A.*, **34**, 827–828.
4. Kearsley,S.K. (1990) An algorithm for the simultaneous superposition of a structural series. *J. Comput. Chem.*, **11**, 1187–1192.
5. Diamond,R. (1992) On the multiple simultaneous superposition of molecular structures by rigid body transformations. *Protein Sci.*, **1**, 1279–1287.
6. Koradi,R., Billeter,M. and Wuthrich,K. (1996) MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.*, **14**, 29–32.
7. Kaplan,W. and Littlejohn,T.G. (2001) Swiss-PDB Viewer (Deep View). *Brief Bioinform.*, **2**, 195–197.
8. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.
9. Willard,L., Ranjan,A., Zhang,H., Monzavi,H., Boyko,R.F., Sykes,B.D. and Wishart,D.S. (2003) VADAR: a web server for quantitative evaluation of protein structure quality. *Nucleic Acids Res.*, **31**, 3316–3319.
10. Richards,F.M. and Kundrot,C.E. (1988) Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. *Proteins.*, **3**, 71–84.
11. Rupp,B. and Parkin. S. (1996) PDBSUP—a FORTRAN program that determines the rotation matrix and translation vector for best fit superposition of two pdb files by solving the quaternion eigenvalue problem. Lawrence Livermore National Laboratory, Livermore, CA. http://www-structure.llnl.gov/xray/comp/superpos.htm
12. Kraulis,P.J. (1991) MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.*, **24**, 946–950.
13. Walther,D. (1997) WebMol — a Java based PDB viewer. *Trends Biochem. Sci.*, **22**, 274–275.
14. Lackner. P., Koppensteiner,W.A., Sippl,M.J. and Domingues,F.S. (2000) ProSup: a refined tool for protein structure alignment. *Protein Eng.*, **13**, 745–752.