Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W111–W115***
doi:10.1093/nar/gki457

AS2TS system for protein structure modeling and analysis

A. Zemla*, C. Ecale Zhou, T. Slezak, T. Kuczmarski, D. Rama, C. Torres, D. Sawicka and D. Barsky

Lawrence Livermore National Laboratory, 7000 East Avenue, Livermore, CA 94550, USA

Received February 15, 2005; Revised and Accepted April 4, 2005

ABSTRACT

We present a set of programs and a website designed to facilitate protein structure comparison and protein structure modeling efforts. Our protein structure analysis and comparison services use the LGA (local-global alignment) program to search for regions of local similarity and to evaluate the level of structural similarity between compared protein structures. To facilitate the homology-based protein structure modeling process, our AL2TS service translates given sequence–structure alignment data into the standard Protein Data Bank (PDB) atom records (coordinates). For a given sequence of amino acids, the AS2TS (amino acid sequence to tertiary structure) system calculates (e.g. using PSI-BLAST PDB analysis) a list of the closest proteins from the PDB, and then a set of draft 3D models is automatically created. Web services are available at http://as2ts.llnl.gov/.

INTRODUCTION

Determination of protein structures via X-ray crystallography or NMR is a relatively slow and expensive process. The difficulty in increasing the rate of experimental determination of protein structures has led to the emphasis on 'computational prediction' and 'analysis' of protein structures. The web page described below has been designed to provide access to several computational protein structure comparison (LGA) and protein structure modeling (AS2TS) services.

PROTEIN STRUCTURE ANALYSIS SERVICES

The ability to verify sequence-based alignments by comparing with the correct structural alignments plays a crucial role in improving the quality of protein structure modeling, protein classification and protein function recognition. The LGA program (1) facilitates this analysis of sequence–structure correspondence. LGA allows detailed pairwise structural comparison of a submitted pair of proteins and also comparison of protein structures or fragments of protein structures with a selected set of proteins from the Protein Data Bank (PDB) (2). The data generated by LGA can be successfully used in a scoring function to rank the level of similarity between compared structures and to allow structural classification when many proteins are being analyzed. LGA also allows the clustering of similar fragments of protein structures. While comparing protein structures, the program generates data that provide detailed information not only about the degree of global similarity but also about regions of local similarity in protein structures. Searching for the best superposition between two structures, LGA calculates the number of residues from the second structure (the target) that are close enough under the specified distance cut-off to the corresponding residues of the first structure (the model). The distance cut-off can be chosen from 0.1 to 10.0 Å in order to calculate a more accurate (tight) or a more relaxed superposition.

There are two provided structural comparison services:

(i) LGA, a protein structure comparison facility, allows the submission of two 3D protein structures or fragments of 3D protein structures (coordinates in the PDB format) for pairwise structural comparative analysis. As a result of LGA processing, a user will receive (a) information about the regions of structural similarity between the submitted proteins and (b) the rotated coordinates of the first structure.

To perform a structural similarity search and to sort the models (templates), the target (i.e. the frame of reference) coordinates can be fixed (placing it as a second structure in all pairwise comparisons). And the user may sort the results (PDB files, models) from LGA processing either by the number of superimposed residues *N* (under the selected distance cut-off), by the GDT_TS score (an average taken from four distance cut-offs), or by the LGA_S structural similarity score [weighted results from the full set of

*To whom correspondence should be addressed at Computing Applications and Research, Energy, Environment, and Biology Division, Lawrence Livermore National Laboratory, 7000 East Avenue, Livermore, CA 94550, USA. Tel: +1 925 423 5571; Fax: +1 925 423 6437; Email: adamz@llnl.gov

distance cut-offs, see (1)]. This multiple pairwise structural comparison is facilitated by the LGA–PDB chain service.

(ii) The LGA–PDB chain structural comparison service allows the submission of a protein structure (target) in the PDB format and a list of selected chains from the list of PDB entries. All chains are structurally compared with the submitted target structure.

Note that when the LGA program is run with options '−1, −2, −3' it does not calculate the structure-based alignments, but calculates only the structural superposition for a given (fixed) residue–residue correspondence. If the user needs to calculate a structural alignment (automatically establish the residue–residue correspondence), then option '−4' should be selected. An explanation and several examples of how to properly select from both structures the desired set of residues for LGA calculations is provided on the website as the service description.

PROTEIN STRUCTURE MODELING SERVICES

The discovery that proteins with even negligible sequence similarity can have similar 3D structures, and can perform similar functions, serves as a foundation for the development of many computational protein structure prediction methods. CASP (3) experiments have shown that protein structure prediction methods based on homology search techniques are still the most reliable prediction methods (4). To facilitate the process of homology-based structural modeling, we have developed a set of services called AS2TS. Provided services are as follows:

(i) The AL2TS [sequence–structure alignment (AL) into tertiary structure (TS)] service is designed to generate a tertiary structure (3D model) for a given sequence–structure alignment model. The alignment model is automatically translated into the TS format in which a given PDB entry

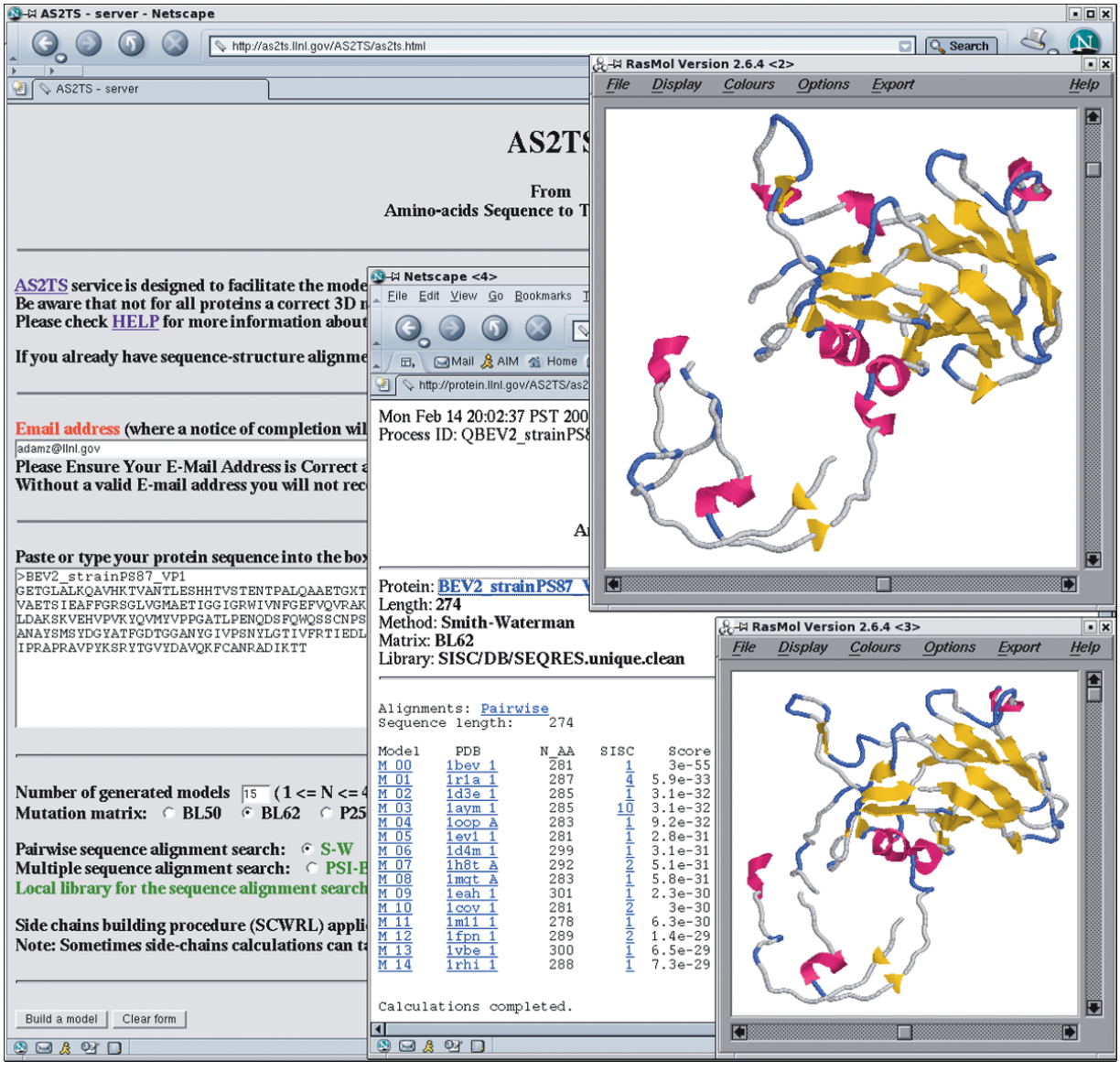


Figure 1. WWW interface to the AS2TS system. For a given sequence of amino acids, the AS2TS system generates a list of the closest templates (related proteins) from the PDB, and creates a set of corresponding 3D models.

is used as a template by which residue coordinates are assigned to corresponding residues in the 3D model. The server accepts any of the three input data formats: AL, which is a standard format used for prediction submissions to CASP experiments, SAL (sequence alignment format) and standard BLAST (5) alignment format.

(ii) The AS2TS [amino acid sequence (AS) into tertiary structure (TS)] service is designed to facilitate the modeling of a tertiary structure (3D model) for a given sequence of amino acids. Using selected sequence alignment search programs, Smith–Waterman (6), FASTA (7), BLAST and PSI-BLAST (5), our AS2TS system searches for homologous proteins in the PDB, calculates alignment models and automatically creates a set of draft 3D models.

(iii) SCWRL is the side chain builder for the AS2TS system. For a given protein structure, SCWRL (Side Chain placement With a Rotamer Library) (8) calculates *de novo* conformation of side chain atoms.

For a given sequence of amino acids, our AS2TS system performs a quick search for the closest PDB homologs that can be used for 3D protein structure modeling. In our system the NR and the PDB data are updated weekly, so generated template information helps the user to estimate the quality of homology-based 3D models that can be currently calculated for a given protein sequence.

Our AS2TS protein structure modeling and analysis system has been used in several collaborative biological research projects (9,10).

EXAMPLES OF RESEARCH IN WHICH OUR SERVICES HAVE BEEN UTILIZED

Models of bovine enterovirus capsid proteins

Figure 1 shows the screenshots of results of using our AS2TS system for a quick search for the closest PDB homologs that could be used for 3D model building of the capsid protein sequences of bovine enterovirus (BEV)-2 strain PS-87. BEVs are members of the Picornaviridae family, genus *Enterovirus*. Detailed 3D protein structure models for three BEV strains were created. This modeling effort was performed in two steps: (i) the structure of the closest template (PDB entry: 1 bev) was modified/corrected in several regions, and some missing residues were modeled; and (ii) the modified 1 bev structure was used as a template to build 3D models for capsid proteins of the three BEV strains of interest.

We have created complete 3D models of the capsids (Figure 2, right) for three BEV strains and for some related PDB templates. Calculated structures will be used for detailed analysis of the 'canyon regions' and for identifying structural differences and similarities among various animal picornaviruses. Modeling of the BEV-2 capsid structure supports the generally accepted idea that the region of the VP-1 protein that connects the eight β-strands making up the wedge-shaped region of each capsid protein is part of the variable region specifying the antigenically variable sites. The details of this work were published previously (9).

Molecular replacement

The AS2TS system has been used to facilitate the molecular replacement (MR) phasing technique in experimental X-ray crystallographic determination of the protein structure of *Mycobacterium tuberculosis* (MTB) RmlC epimerase (Rv3465) from the strain H37rv. The MTB RmlC protein was crystallized by the Biosciences crystallography group at Lawrence Livermore National Laboratory, and native X-ray data (without phases) were collected at the Advanced Light Source at Lawrence Berkeley Laboratory. Although structurally related homologs were tried for MR, the technique failed because the sequences were too dissimilar. Using our AS2TS system, we built two homology models of this protein that were then successfully employed as MR targets (10).

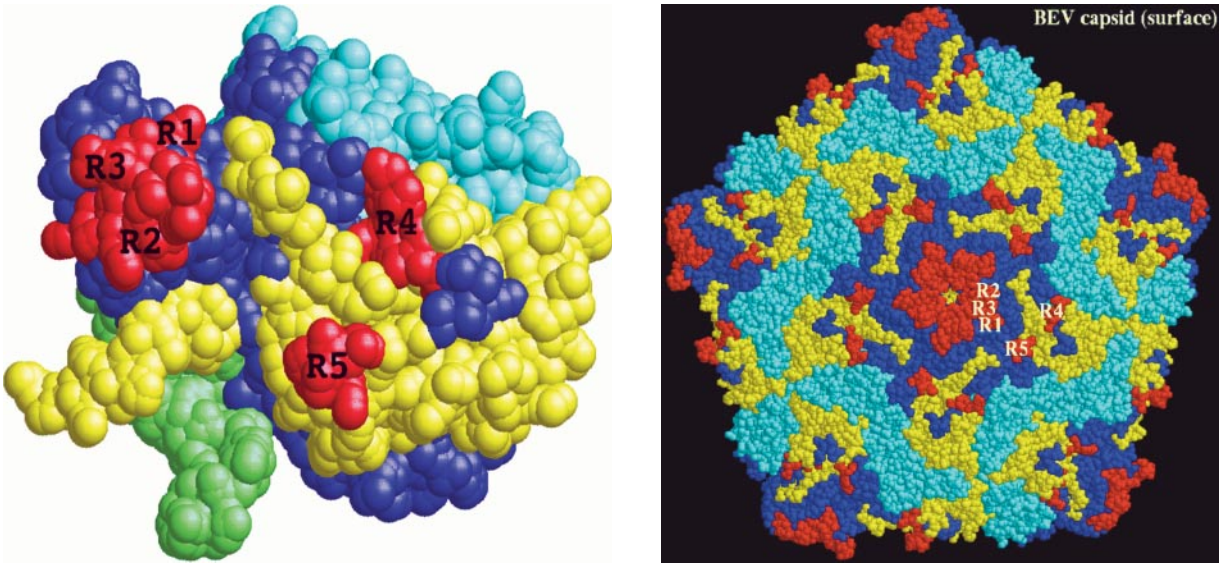
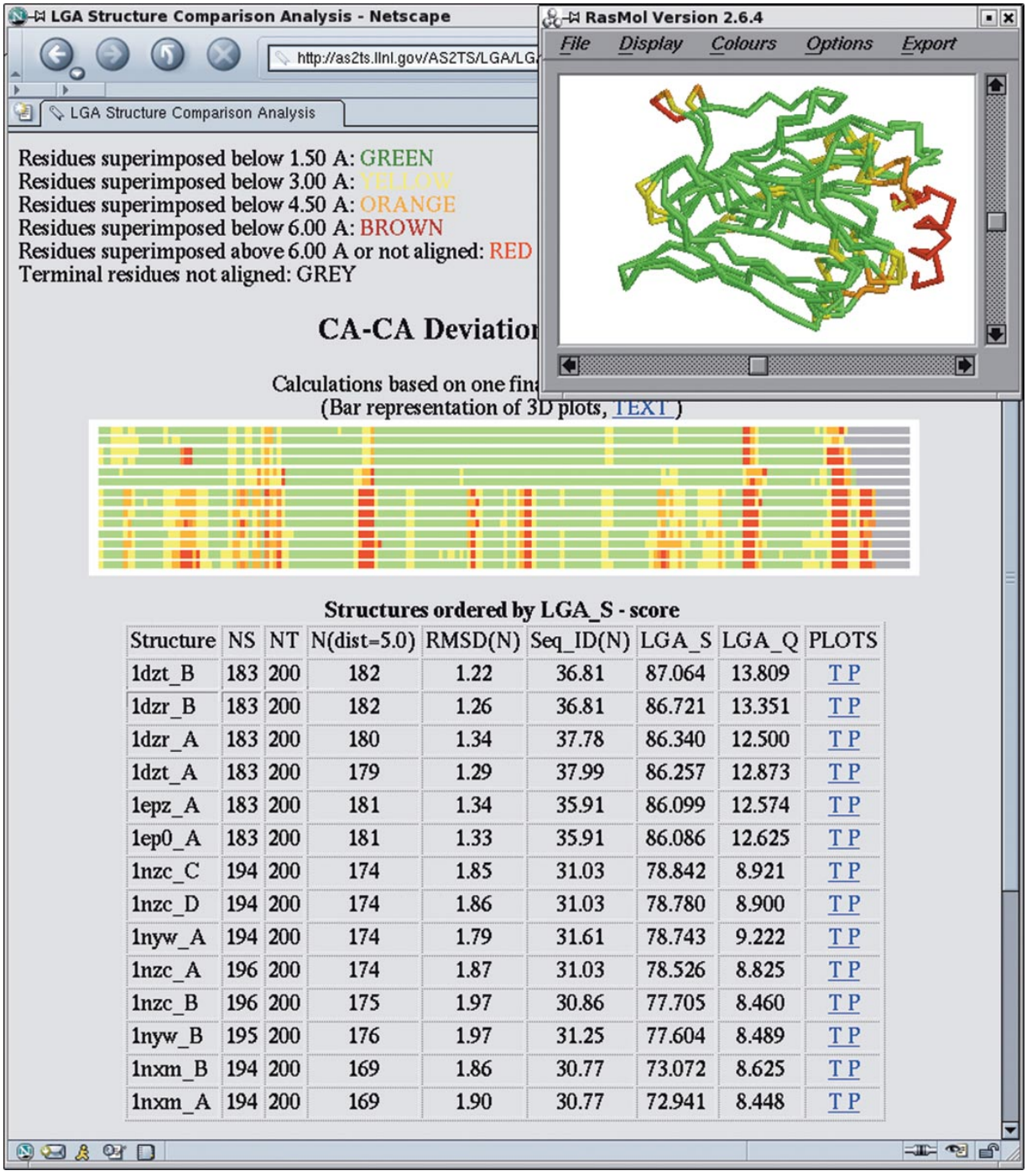


Figure 2. A 3D model (left) of the protomer for BEV-2 strain PS-87. Protein VP-1 is in blue, VP-2 in cyan, VP-3 in yellow and VP-4 in green. Insertion and deletion regions R1–R5 are in red. The surface (right) of the BEV capsid contains 15 assembled protomers. Protein VP-4 is not visible on the capsid plot because it is completely buried under the surface. Insertion and deletion regions R1–R3 and R4 are located on the rims of the observed canyon with R5 lying in its base [from (9)].



Structure	NS	NT	N(dist=5.0)	RMSD(N)	Seq_ID(N)	LGA_S	LGA_Q	PLOTS
1dzt_B	183	200	182	1.22	36.81	87.064	13.809	T P
1dzr_B	183	200	182	1.26	36.81	86.721	13.351	T P
1dzr_A	183	200	180	1.34	37.78	86.340	12.500	T P
1dzt_A	183	200	179	1.29	37.99	86.257	12.873	T P
1epz_A	183	200	181	1.34	35.91	86.099	12.574	T P
1ep0_A	183	200	181	1.33	35.91	86.086	12.625	T P
1nzc_C	194	200	174	1.85	31.03	78.842	8.921	T P
1nzc_D	194	200	174	1.86	31.03	78.780	8.900	T P
1nyw_A	194	200	174	1.79	31.61	78.743	9.222	T P
1nzc_A	196	200	174	1.87	31.03	78.526	8.825	T P
1nzc_B	196	200	175	1.97	30.86	77.705	8.460	T P
1nyw_B	195	200	176	1.97	31.25	77.604	8.489	T P
1nxm_B	194	200	169	1.86	30.77	73.072	8.625	T P
1nxm_A	194	200	169	1.90	30.77	72.941	8.448	T P

Figure 3. Pairwise structural alignment of 14 homologous proteins with RmlC from the MTB using the LGA–PDB chains comparison service. Colored bars represent Cα–Cα distance deviation between superimposed PDB structures and RmlC [200 residues; from the left (N-terminal) to the right (C-terminal)]. Residues superimposed <1.5 Å are in green, <3.0 Å in yellow, <4.5 Å in orange, <6.0 Å in brown and residues ⩾6.0 Å in red. Not aligned terminal residues are in gray. The table below the bars contains information (in the same order as bars) about the level of sequence identity (Seq_ID), level of structural similarity (LGA_S) and r.m.s.d. in Å calculated on all Cα pairs that are superimposed under 5 Å distance cut-off. For example, this plot shows that all the homologous proteins differ significantly (red) from RmlC in the C-terminal part (loop 160–165, region 179–186), and also that the C-terminal helix is not present (gray) in all the templates. The rasmol plot shown on top represents the first bar (superposition between RmlC and 1dzt_B structures).

Evaluation of the generated MTB RmlC models was performed using LGA. Detailed structural comparison analysis of 14 homologs revealed two proteins, dTDP-4-dehydrorhamnose epimerase (PDB entry: 1ep0) and RmlC from *Salmonella typhimurium* (PDB entry: 1dzr), which were selected as primary templates.

Figure 3 illustrates the results from LGA analysis when 14 proteins of known structure were compared with the selected

target protein. This LGA capability allowed us to localize the regions that were structurally similar among all analyzed proteins, select one or more structures as a template(s) for homology modeling, and use this information to create a consensus model. The process of structural determination for the MTB RmlC protein (PDB entry: 1upi) was described by Kanterdjieff *et al*. (10).

ACKNOWLEDGEMENTS

This work was performed under the auspices of the US Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48. The design and development of described systems was supported by LLNL LDRD grants 02-LW-003 and 04-ERD-068 to A.Z. Funding to pay the Open Access publication charges for this article was provided by US Department of Energy.

Conflict of interest statement. None declared.

REFERENCES

1. Zemla,A. (2003) LGA—a method for finding 3D similarities in protein structures. *Nucleic Acids Res*., **31**, 3370–3374.
2. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*., **28**, 235–242.
3. Moult,J., Fidelis,K., Zemla,A. and Hubbard,T. (2003) Critical assessment of methods of protein structure prediction (CASP)-round V. *Proteins*, **S6**, 334–339.
4. Venclovas,C., Zemla,A., Fidelis,K. and Moult,J. (2003) Assessment of progress over the CASP experiments. *Proteins*, **S6**, 585–595.
5. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.
6. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol*., **147**, 195–197.
7. Pearson,W.R. (1991) Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith–Waterman and FASTA algorithms. *Genomics*, **11**, 635–650.
8. Bower,M., Cohen,F.E. and Dunbrack,R.L.,Jr (1997) Sidechain prediction from a backbone-dependent rotamer library: a new tool for homology modeling. *J. Mol. Biol*., **267**, 1268–1282.
9. Goens,S.D., Botero,S., Zemla,A., Ecale Zhou,C. and Perdue,M. (2004) Bovine enterovirus type 2. Complete genomic sequence and molecular modeling of the reference strain and a wild type isolate from endemically infected US cattle. *J. Gen. Virol*., **85**, 3195–3203.
10. Kanterdjieff,K.A., Kim,Ch.Y., Naranjo,C., Waldo,G.S., Lekin,T.P., Segelke,B.W., Zemla,A., Park,M.S., Terwilliger,T.C. and Rupp,B. (2004) *Mycobacterium tuberculosis* RmlC epimerase (Rv3465): a promising drug-target structure in the rhamnose pathway. *Acta Crystallogr. D*, **60**, 895–902.