W606–W609 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh400

CaspR: a web server for automated molecular replacement using homology modelling

Jean-Baptiste Claude, Karsten Suhre*, Cédric Notredame, Jean-Michel Claverie and Chantal Abergel

Information Génomique & Structurale (UPR CNRS 2589), Institut de Biologie Structurale et Microbiologie, 31, chemin Joseph Aiguier, 13402 Marseille Cedex 20, France

Received February 13, 2004; Revised March 15, 2004; Accepted March 25, 2004

ABSTRACT

Molecular replacement (MR) is the method of choice for X-ray crystallography structure determination when structural homologues are available in the Protein Data Bank (PDB). Although the success rate of MR decreases sharply when the sequence similarity between template and target proteins drops below 35% identical residues, it has been found that screening for MR solutions with a large number of different homology models may still produce a suitable solution where the original template failed. Here we present the web tool CaspR, implementing such a strategy in an automated manner. On input of experimental diffraction data, of the corresponding target sequence and of one or several potential templates, CaspR executes an optimized molecular replacement procedure using a combination of well-established stand-alone software tools. The protocol of model building and screening begins with the generation of multiple structure–sequence alignments produced with T-COFFEE, followed by homology model building using MODELLER, molecular replacement with AMoRe and model refinement based on CNS. As a result, CaspR provides a progress report in the form of hierarchically organized summary sheets that describe the different stages of the computation with an increasing level of detail. For the 10 highest-scoring potential solutions, pre-refined structures are made available for download in PDB format. Results already obtained with CaspR and reported on the web server suggest that such a strategy significantly increases the fraction of protein structures which may be solved by MR. Moreover, even in situations where standard MR yields a solution, pre-refined homology models produced by CaspR significantly reduce the time-consuming refinement process. We expect this automated procedure to have a significant impact on the throughput of large-scale structural genomics projects. CaspR is freely available at http://igs-server.cnrs-mrs.fr/Caspr/.

INTRODUCTION

Molecular replacement (MR) is the most cost-effective method for solving the three-dimensional (3D) structure of a protein by X-ray crystallography. However, the MR approach requires the availability of at least one close structural homologue. Thanks to the ongoing structural genomics projects, the Protein Data Bank (PDB) (1) is now rapidly growing, increasing the probability of finding structural homologues. At the same time, bioinformatics techniques for detecting low sequence similarity keep improving, allowing more distant putative 3D homologues to be identified. MR is thus expected to play an increasing role in the phasing of protein X-ray diffraction data. In most cases of successful MR application, the sequences of the protein of interest and of the structural homologue are at least 35% identical. Below that threshold, and down to 20% of identical residues, the overall fold is usually well conserved but the differences in the 3D structures become too large to be handled by the standard MR protocol. Homology modelling has been proposed (2) to extend the application of MR to these cases of lower sequence similarity. An example of such a procedure is already implemented [MODELLER, (3)] in the CCP4 software package to improve the initial model after MR solutions have been found (4).

Here we present an automated protocol based on two main principles. First, sequence and structural information are combined using a new multiple-alignment program (5) to generate higher-quality homology models. Second, a large number of different models are screened for MR solutions. The implementation of this protocol in the CaspR web server includes

*To whom correspondence should be addressed. Tel: +33491164604; Fax: +33491164549; Email: karsten.suhre@igs.cnrs-mrs.fr

the automatic excision of unreliably aligned residues from the 3D models. This protocol was successfully applied (K. Suhre, manuscript in preparation) to solve the crystal structures of three *Escherichia coli* proteins of unknown function in the context of the structural genomics project BIGS [(6), http://igs-server.cnrs-mrs.fr/BIGS/]. These proteins are (i) YecD (four molecules/a.u.) sharing less than 25% sequence identity with two known structures, (ii) YggV (two molecules/a.u.) with 33% identity with one related structure and (iii) YahK (one molecule/a.u.) with 32% sequence identity to three known structures. In all the above cases standard MR protocols failed to identify a solution using the available structural homologues, while models generated through the CaspR procedure provided a convergent solution up to the final refinement step. These three cases are used as walk-through examples on our web server.

IMPLEMENTATION

The CaspR web server is built around a set of standard software tools widely used within the protein crystallography and bioinformatics communities. The first step in the process (see also Figure 1) is to produce a reliable multiple alignment using the T-COFFEE software (7). A specific feature of T-COFFEE is to provide a reliability index [CORE index, (8)] for each

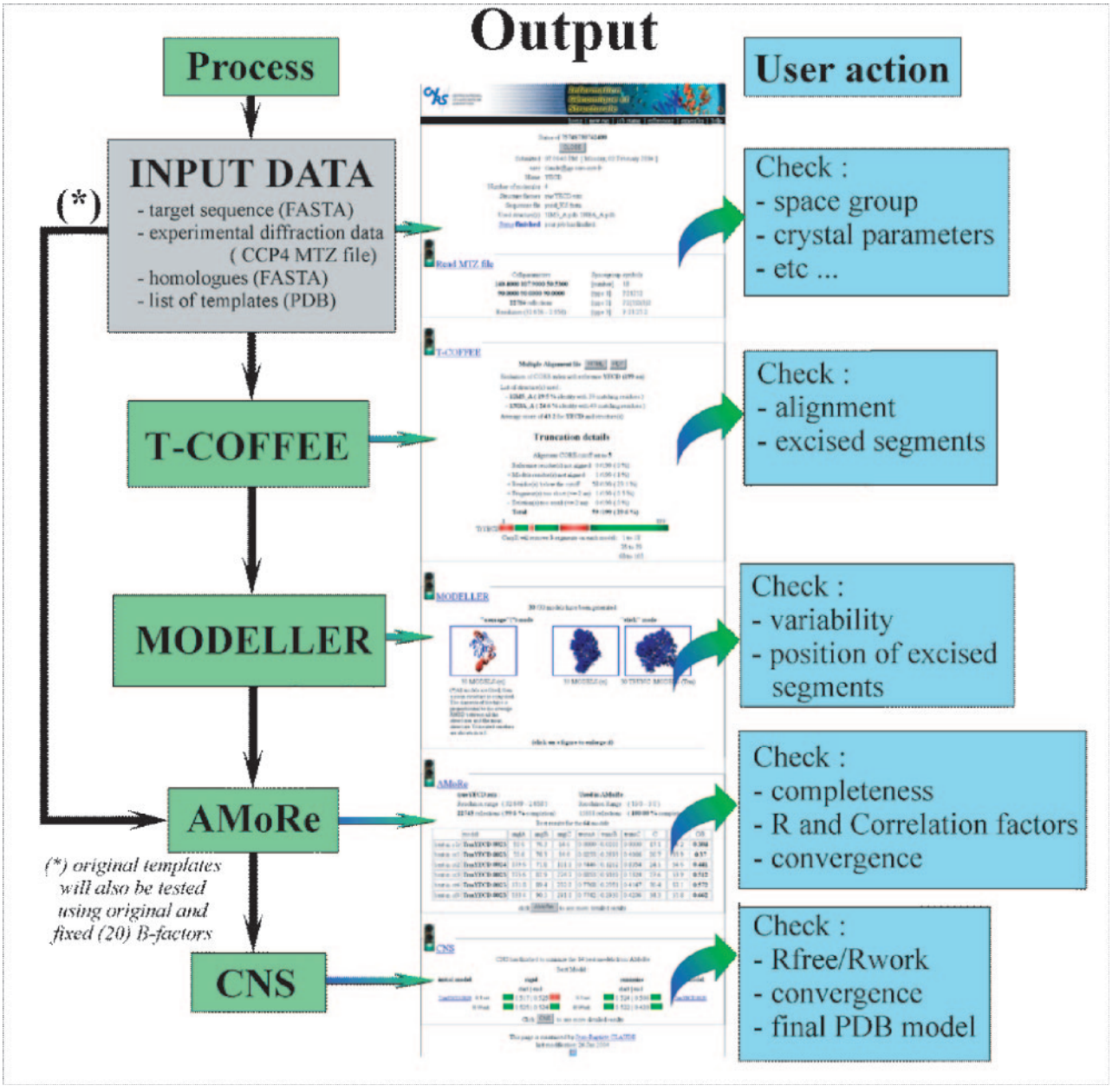

Figure 1. The process of automatic MR using screening with homology models as implemented in CaspR. Typical CaspR outputs are shown as screenshots. Actions to be taken by the user to verify the proceeding of a CaspR run are presented on the right. Final validation of the CaspR solution(s) will of course be provided by the observation of the electronic density maps computed and analysed by the user.

position in the alignment. Based on the value of this index, suitable segments of the target sequence/structure are identified and unreliably aligned segments are automatically excised. In its most recent version (3D-COFFEE; 5), T-COFFEE combines structure and sequence information to generate better multiple alignments and in turn improves the quality of the homology models produced by MODELLER. Another key feature of our approach is the screening of a large number of these 3D models, all being generated from a moderately perturbed starting model (2). These models are automatically screened in search of a molecular replacement solution using the AMoRe software (9). In a final step, the solutions are automatically pre-refined using CNS (10–12), where the convergence of the free and working R-factors is the final criterion for the ranking of the different solutions.

USING THE WEB SERVER

A job submission to the CaspR web server must include (i) the target protein sequence file (in FASTA format), optionally, one or several additional sequences of homologous proteins (used to optimize the alignment process) may be appended to this file; (ii) the crystallographic structure factors in truncated MTZ format (as in CCP4); (iii) the PDB identifiers of one or more structural neighbours; (iv) auxiliary crystallographic information (i.e. expected number of molecules per asymmetric unit) and (v) an email address.

A typical CaspR run takes between 2 and 48 h, depending on the protein size, the space group, number of molecules in the asymmetric unit and the server load. The job status can be monitored on the CaspR web server. Submitted data will be kept confidential and can be removed any time by the user, while logfiles will be used for further optimization of the CaspR process.

The organization of CaspR output is presented in Figure 1, as well as the various controls to be performed by the user at the different stages of the process. In addition, active links are provided for the display of the T-COFFEE outputs, of the PDB co-ordinates of all the models submitted to AMoRe and of the AMoRe statistics, as well as the 10 best-ranking structures.

RESULTS

Four test cases using data retrieved from the PDB and three cases corresponding to experimental data produced in our laboratory (http://igs-server.cnrs-mrs.fr/BIGS) have been used to validate the CaspR suite through different MR problems of various levels of complexity. Details are available as supplementary material on the CaspR web site, together with the complete results of the CaspR runs (logfiles). In summary, easy MR cases (1MP0 using 1N8K as a template, YhbO) are easily solved using the CaspR procedure, and the proposed models always exhibit better R-factors than the original template after CNS refinement. In five other cases (1AJX, 1K6K using 1M3E as a template, YahK, YggV, YecD) the original templates do not produce a valid MR solution using a standard procedure whereas CaspR succeeds in finding a converging MR solution. Among them, YecD (PDB id 1J2R) is the first occurrence of a structure uniquely solved using the CaspR procedure. Finally, there are two cases (1MP0 using 1JVB and 1JQB, 1K6D using 1POI) that remain presently unsolved by MR, and are thus a good benchmark for future improvements of our procedure.

CONCLUDING REMARKS

By using structure and sequence information together to generate homology models the CaspR web server is pushing back the limits of structure solving using MR. Its purpose is to provide the structural genomics community with a powerful tool that is expected to reduce the need for expensive and time-consuming phasing experiments such as MIR and MAD. CaspR is also useful in simple MR cases by automatically replacing the amino acid sequence of the template by those of the molecule of interest, thus accelerating the tedious refinement process.

In difficult cases, the limiting factor is the information content of the multiple alignment used to link the target protein sequence to the available structure(s), and thus to generate the models. A standardization (and optimization) of this step might be achieved by limiting the user input to the sequence of the crystallized protein and the experimental diffraction data. CaspR would then automatically identify the proper sequence subset to be used, i.e. the one providing the most gradual evolutionary transition from the target to the structural template. Along the same lines, the Molecular Modeling Database [MMDB, (13)] can also be used to optimize the selection of the best structural representatives within a given family. Finally, the CaspR web server will eventually be installed on a large cluster of Linux machines (and/or run on a grid) to reduce its computing time and adapt its performance to the needs of the structural genomics community by allowing a large number of jobs to be run in parallel.

ACKNOWLEDGEMENTS

Protein models used as templates in CaspR are continuously updated from the PDB (1). We gratefully acknowledge the use of the software tools included in CaspR: CNS (10); LSQMAN from the DéjàVu package (14); AMoRe (9); T-COFFEE (7); MODELLER (3) and MOLMOL (15).

REFERENCES

1. 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.
2. Jones,D.T. (2001) Evaluating the potential of using fold-recognition models for molecular replacement. *Acta Cryst.*, **D57**, 1428–1434.
3. Sali,A. and Blundell,T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.
4. Collaborative Computational Project (2002) High-throughput structure determination. Proceedings of the 2002 CCP4 study weekend. *Acta Cryst.*, **D58**, 1897–1970.
5. Poirot,O., Suhre,K., Abergel,C., O'Toole,E. and Notredame,C. (2004) 3DCoffee@igs: a web server for combining sequences and structures into a multiple sequence alignment. *Nucleic Acids Res.*, **32**, W37–W40.
6. Abergel,C., Coutard,B., Byrne,D., Chenivesse,S., Claude,J.-B., Deregnaucourt,C., Fricaux,T., Boutreux,C., Jeudy,S., Lebrun,R., Maza,C., Notredame,C., Poirot,O., Suhre,K., Varagnol,M. and Claverie,J.-M. (2003) Structural genomics of highly conserved microbial genes of unknown function in search of new antibacterial targets. *J. Struct. Funct. Genomics*, **4**, 141–157.

7. Notredame,C., Higgins,D. and Heringa,J. (2000), T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.
8. Notredame,C. and Abergel,C. (2003) Using T-Coffee to assess the reliability of multiple sequence alignments. In: Andrade,M.A. (ed.), *Bioinformatics and Genomes*, Horizon Scientific Press, Wymondham, UK, pp. 27–49.
9. Navaza,J. (2001) Implementation of molecular replacement in AMoRe. *Acta Cryst.*, **D57**, 1367–1372.
10. Brünger,A.T., Adams,P.D., Clore,G.M., DeLano,W.L., Gros,P., Grosse-Kunstleve,R.W., Jiang,J.-S., Kuszewski,J., Nilges,M., Pannu,N.S., Read,R.J., Rice,L.M., Simonson,T. and Warren,G.L. (1998) Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Cryst.*, **D54**, 905–921.
11. Pannu,N.S. and Read,R.J. (1996) Improved structure refinement through maximum likelihood. *Acta Cryst.*, **A52**, 659–668.
12. Adams,P.D., Pannu,N.S., Read,R.J. and Brunger,A.T. (1997) Cross-validated maximum likelihood enhances crystallographic simulated annealing refinement. *Proc. Natl Acad. Sci. USA*, **94**, 5018–5023.
13. Chen,J., Anderson,J.B., DeWeese-Scott,C., Fedorova,N.D., Geer,L.Y., He,S., Hurwitz,D.I., Jackson,J.D., Jacobs,A.R., Lanczycki,C.J., Liebert,C.A., Liu,C., Madej,T., Marchler-Bauer,A., Marchler,G.H., Mazumder,R., Nikolskaya,A.N., Rao,B.S., Panchenko,A.R., Shoemaker,B.A., Simonyan,V., Song,J.S., Thiessen,P.A., Vasudevan,S., Wang,Y., Yamashita,R.A., Yin,J.J. and Bryant,S.H. MMDB: Entrez's 3D-structure database. *Nucleic Acids Res.*, **31**, 474–477.
14. Kleywegt,G.J. (1996) Use of non-crystallographic symmetry in protein structure refinement. *Acta Cryst.*, **D52**, 842–857.
15. Koradi,R., Billeter,M. and Wüthrich,K. (1996) MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graphics*, **14**, 51–55.