

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

**Jean-Baptiste Claude, Karsten Suhre<sup>\*</sup>, Cédric Notredame, Jean-Michel Claverie & Chantal Abergel**

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

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

**KEYWORDS** : molecular replacement, structural genomics, X-ray crystallography, homology modelling, multiple alignment

### **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 decrease 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 (Notredame *et al.*, *J. Mol. Biol.*, 2000), followed by homology model-building using MODELLER (Sali & Blundell, *J. Mol. Biol.*, 1993), molecular replacement with AMoRe (Navaza, *Acta Cryst.*, 2001), and model refinement based on CNS (Brunger, *Acta Cryst.*, 1998). As a result, CaspR provides a progress report in form of hierarchically organized summary sheets that describe the different stages of the computation with an increasing level of detail. For the ten 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 may 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 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 structural homologues to be found. At the same time, bioinformatic techniques for detecting low sequence similarity keep improving, allowing more distant putative 3-D 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 sequence of the protein of interest and of the structural homologue were 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 3-D structures become too large to be handled by 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 (Poirot et al., submitted) 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 the automatic excision of unreliably aligned residues from the 3D models. This protocol was successfully applied (Suhre et al., in preparation) to solve the crystal structures of three *E. coli* proteins of unknown function in the context of the structural genomics project BIGS (5, <http://igs-server.cnrs-mrs.fr/BIGS/>). These proteins are 1) *YecD* (4 molecules/a.u.) sharing less than 25% sequence identity with two known structures, 2) *YggV* (two molecules/a.u.) with 33% identity with one related structure, and 3) *YahK* (one molecule/a.u.) with 32% sequence identity to three known structures. In all of 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 Fig. 1) is to produce a reliable multiple alignment using the T-COFFEE software (6). A specific feature of T-COFFEE is to provide a reliability index (CORE index, 7) for each 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; Poirot et al., submitted), 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 for a molecular replacement solution using the AMoRe software (8). In a final step, the solutions are automatically pre-refined using CNS (9 - 11), where the convergence of the free and working R-factors is the final criteria for the ranking of the different solutions.

## **USING THE WEB SERVER**

A job submission to the CaspR web-server must include

- 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,
- the crystallographic structure factors in truncated MTZ format (as in CCP4),
- the PDB-identifiers of one or more structural neighbours,
- auxiliary crystallographic information (i.e. expected number of molecules per asymmetric unit)
- an E-mail address

A typical CaspR run takes between 2 to 48 hours, depending of 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 optimisation of the CaspR process.

The organization of CaspR output is presented in Fig. 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 coordinates of all the models submitted to AMoRe, 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 (log-files). 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 while 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 optimisation) 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 line, the Molecular Modeling Database (MMDB, 12) can also be used to optimise 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

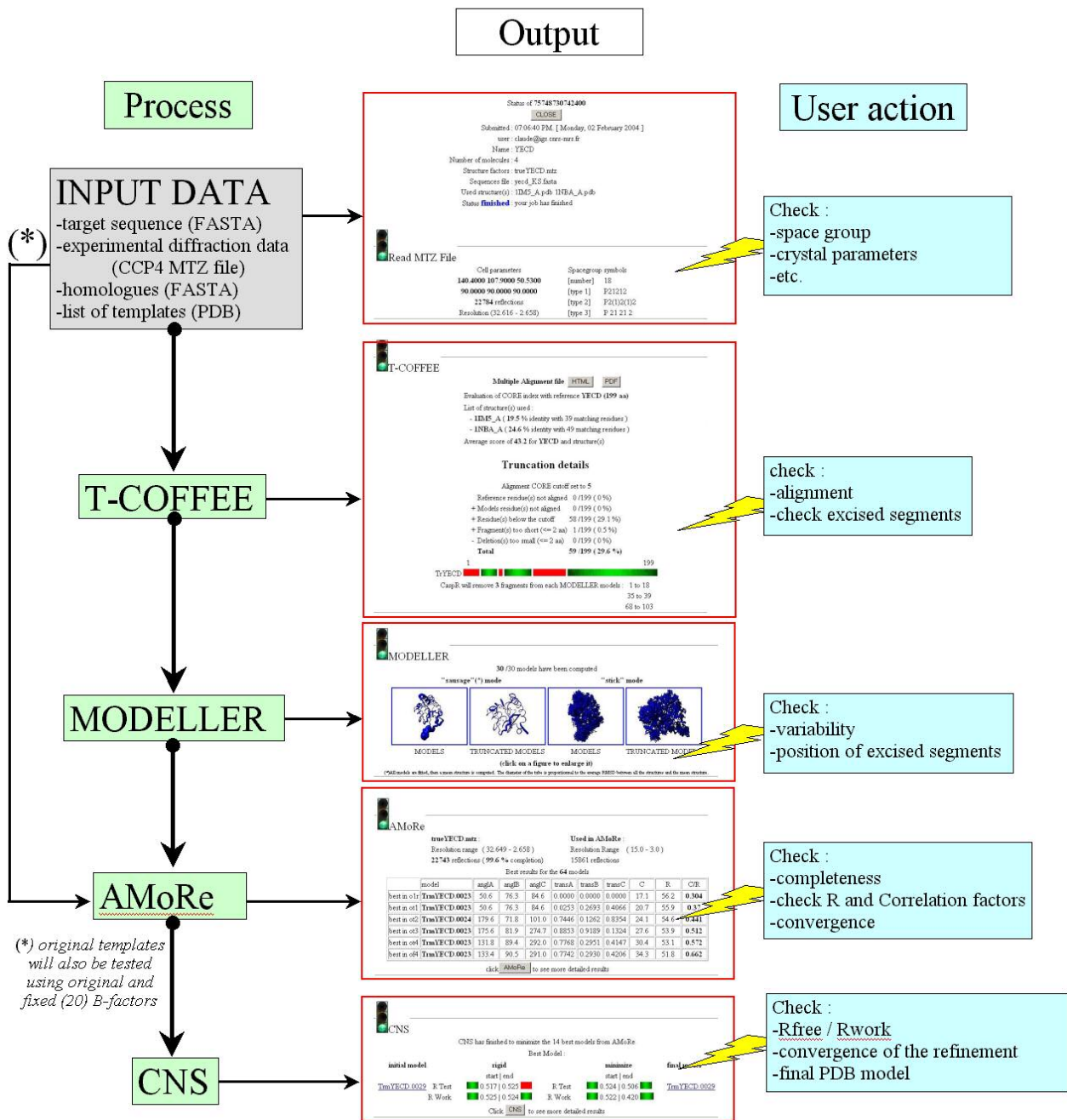
Protein models used as templates in CaspR are continuously updated from the PDB (1). We gratefully acknowledge the use of the software tools included in CaspR: CNS (9); LSQMAN from the DéjàVu package (13); AMoRe (8); T-COFFEE (6); MODELLER (3) and MOLMOL (14).

## REFERENCES

1. Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N. & 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. & 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. Abergel C., Coutard B., Byrne D., Chenivresse S., Claude J-B., Deregnaucourt C., Fricaux T., Boutreux C., Jeudy S., Lebrun R., Maza C., Notredame C., Poirot O., Suhre K., Varagnol M. & 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.
6. Notredame, C., Higgins, D. & Heringa, J. (2000), T-Coffee: A novel method for fast and accurate multiple sequence alignment, *J. Mol. Biol.* **302**, 205-217.
7. Notredame C. and Abergel C. (2003) Using T-Coffee to assess the reliability of multiple sequence alignments. *Bioinformatics and Genomes (M.A. Andrade, ed.) Horizon Scientific Press, Wymondham, UK*, pp 27-49.
8. Navaza, J. (2001) Implementation of molecular replacement in AMoRe. *Acta Cryst.* **D57**, 1367-1372.
9. 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 NS, Read RJ, Rice LM, Simonson T, Warren GL (1998) Crystallography & NMR system: A new software suite for macromolecular structure determination , *Acta Cryst.* **D54**, 905-921.

10. Pannu N.S. and Read R.J. (1996) Improved Structure Refinement Through Maximum Likelihood *Acta Cryst.* **A52**, 659-668.
11. Adams PD, Pannu NS, Read RJ, Brunger AT (1997) Cross-validated maximum likelihood enhances crystallographic simulated annealing refinement. *Proc Natl Acad Sci U S A.* **94**, 5018-5023.
12. Chen J, Anderson JB, DeWeese-Scott C, Fedorova ND, Geer LY, He S, Hurwitz DI, Jackson JD, Jacobs AR, Lanczycki CJ, Liebert CA, Liu C, Madej T, Marchler-Bauer A, Marchler GH, Mazumder R, Nikolskaya AN, Rao BS, Panchenko AR, Shoemaker BA, Simonyan V, Song JS, Thiessen PA, Vasudevan S, Wang Y, Yamashita RA, Yin JJ, Bryant SH. MMDB: Entrez's 3D-structure database. *Nucleic Acids Res.* 2003 Jan 1;31(1):474-7.
13. Kleywegt, G. J. (1996) Use of non-crystallographic symmetry in protein structure refinement. *Acta Cryst.*, **D52**, 842-857.
14. Koradi, R., Billeter, M., and Wüthrich, K. (1996) MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graphics*, **14**, 51-55.

# FIGURES



**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

## **SUPPLEMENTARY MATERIAL**

The following text will be made available as supplementary material on the CaspR web-site. It corresponds to results to be discussed in our research paper (Suhre et al., in preparation) on the method implemented in CaspR. It describes the details of the different cases that were tested on the CaspR web-server (the corresponding log-files and results are available on the site for examination by the reviewers).

### **1AJX: the HIV<sub>1</sub> protease**

The HIV<sub>1</sub> protease structure belongs to the acid proteases fold. It is a homodimer of a 99 residue long protein and can adopt two conformations upon ligand binding (RMSD between 1.0 and 1.2 Å over 99 residues based on C $\alpha$  superimposition of the two structural conformations). Each of these structures has been solved independently and correspond to 1AJX (1) and 1HHP (2) PDB accession numbers. We used the available 1HHP structure to solve the 1AJX structure with CaspR. In this example, the crystals belong to the orthorhombic space group P21212 and there are 2 molecules in the asymmetric unit. While the 1HHP structure clearly identifies the first position in the AMoRe step of the CaspR process, it fails to identify the second molecule using the standard procedure. On the other hand, using the CaspR automated protocol, 10 out of the 30 models produce a two molecule solution in AMoRe and converge during the refinement process (best model R<sub>free</sub>: 38.4, R<sub>work</sub>: 31.4; 1HHP R<sub>free</sub>: 54.8, R<sub>work</sub>: 50.6).

### **1K6D : the E. coli $\alpha$ subunit of the acetate CoA-transferase**

The 1K6D structure (3) belongs to the CoA transferase fold. It crystallizes in the hexagonal space group P62 with two molecules in the asymmetric unit. This 220 residues long protein has a closely related structural homologue exhibiting 35% sequence identity over its entire length (1M3E PDB accession number) as well as a more distant homologue with less than 25% sequence identity (1POI PDB accession number). Both potential templates were tested using the CaspR web-server. In the case of 1M3E, 30 models were produced by MODELLER and residues that were poorly aligned to the target were automatically excised based on the T-COFFEE core index (residues 1 to 7, 124 to 134 and 216 to 220). As a result, the 14 best AMoRe solutions were obtained using the truncated models. 28 out of the 30 truncated and 21 out of the 30 non truncated ones produced an AMoRe solution for both molecules in the a.u.. All of them present a higher correlation coefficient and a lower R-factor in AMoRe than the unperturbed 1M3E structure. The 10 best solutions all correspond to the truncated models and converge during the pre-refinement procedure using CNS (best solution R<sub>free</sub>: 43.7, R<sub>work</sub>: 36.3; 1M3E: R<sub>free</sub>: 56.9 R<sub>work</sub>: 53.2)

In the second case (1POI), 60 models were generated, 30 of them corresponding to the truncated forms based on the T-COFFEE core index (excision of residues 58 to 60 and 218 to 220). In this case, the best AMoRe solution is given by the 1POI structure itself, while none of the solutions converged during the pre-refinement process. This is a case where CaspR was not able to improve the solution with a homology model. Neither the models nor the related structure could produce a result in this standard procedure. A way to retrieve a solution may be to change the core index threshold value to be used in the design of truncated models.

### **1MP0: The human glutathione-dependent formaldehyde dehydrogenase**

The 1MP0 structure (4) is a two domains structure belonging to the GroES-like fold for the first domain and to the NAD(P)-binding Rossmann-fold for the second domain. The crystals belong to the tetragonal space group P43212 and there are 2 molecules of this 373 residues long protein in the asymmetric unit. A closely related structure (1N8K) sharing 56% sequence identity over 373 residues as well as more distantly related one (1JQB and 1JVB) sharing less than 26% sequence identity with 1MP0 were tested through the CaspR web-server.

In the first case (1N8K), 60 models were produced by MODELLER with 30 truncated models based on the T-COFFEE core index (excision of residues 1 to 5 and 245 to 251). The best solution in AMoRe correspond to the 1N8K structure itself. However, all truncated and non truncated models also produced a solution which converged through the pre-refinement procedure using CNS. Eventually, the 10 best solutions in CNS all correspond to homology models and score higher than 1N8K itself, most likely because the models carry the side chains that are proper to 1MP0 (best solution  $R_{\text{free}}$ : 41.4,  $R_{\text{work}}$ : 35.4; 1N8K:  $R_{\text{free}}$ : 48.8  $R_{\text{work}}$ : 49.1. This shows that even in “easy” MR cases, CaspR is still helpful by improving the starting model prior to manual refinement.

In the second case, 60 models were produced by MODELLER based on the two 1JVB and 1JQB structures. The 30 truncated models correspond to the N-terminal removal of the 4 first residues of the 1MP0 sequence. Both the probe structures and the models failed to produce a convincing solution in AMoRe and did not converge in the CNS refinement procedure. As for 1K6D, the lowering of the T-COFFEE core index threshold may be a way to generate new truncated models for solution screening in AMoRE. Molecular replacement can also be run using the two sub-domains independently for the model generation and the two-body screening for solutions in the AMoRe step.

### **BIGS protein YhbO (1OI4): The *E. coli* potential cysteine peptidase protein**

YhbO corresponds to a 193 residues long protein annotated as a potential cysteine peptidase (Swiss-Prot: P45470). It belongs to the Flavodoxin-like fold and shares 41% identity with its structural homologue 1G2I. The crystals belongs to the orthorhombic space group P212121 with 2 molecules in the asymmetric unit. CaspR produced 60 models corresponding to 30 truncated forms (excision of residues 1 to 24; 56 to 63 and 168 to 171) which went through the molecular replacement procedure using AMoRe. The best scoring AMoRe solution is given by the 1G2I structure itself. However all 60 models also produce a correct solution. After refinement procedure with CNS, the 1G2I solution eventually only ranks third below two truncated models with a better convergence, which again can be explained by the replacement of the 1G2I sequence by the proper YhbO one (best solution  $R_{\text{free}}$ : 44.9,  $R_{\text{work}}$ : 37.5; 1G2I:  $R_{\text{free}}$ : 45.7  $R_{\text{work}}$ : 38.4). This again is a case where CaspR is helpful to decrease the time spent in refinement procedure.

### **BIGS protein YahK (1UUF): The *E. coli* Zinc-type alcohol dehydrogenase-like protein**

YahK is a 349 residue long protein annotated as a Zinc-type alcohol dehydrogenase-like protein (Swiss-Prot: P75691). Like 1MP0, it is a two domain structure belonging to the GroES-like fold for the first domain and to the NAD(P)-binding Rossmann-fold for the second domain. It shares 33% identity with its closest homologue 1LLU and 27.6% and 26.5% identity with the two related structures 1H2B and 1JVB. The crystals belongs to the monoclinic space group C2 and there is one molecule in the asymmetric unit. Out of the 60 models produced using the 3 template structures through the CaspR process, 30 correspond to truncated forms based on the T-COFFEE core index (10 excised segments corresponding to positions 1 to 48, 72 to 80, 110 to 115, 129 to 146, 156 to 168, 191 to 193, 248 to 260, 290 to 304, 327 to 336 and 364 to 369). Both 1LLU and 1JVB produces higher scoring solution in the AMoRe search, however, the best solution in the pre-refinement procedure is produced by one of the 30 non truncated models (best solution  $R_{\text{free}}$ : 49.9,  $R_{\text{work}}$ : 40.7; 1LLU:  $R_{\text{free}}$ : 51.1  $R_{\text{work}}$ : 42.0). Due to their ranking, none of the truncated models went through the refinement process.

### **BIGS protein YggV (also available as 1K7K) : The *E. coli* HAM1 NTPase protein**

YggV is a 197 residue long protein belonging to the HAM1 family of pyrophosphatase (Swiss-Prot: P52061). According to the SCOP database it belongs to the Anticodon-binding domain-like and shares 33.5% identity with the related 2MJP structure. The crystals belong to the tetragonal space group P43212 and there is one molecule in the asymmetric unit. 30 full length models having been produced through the automated procedure and generated 30 truncated models based on T-

COFFEE core index (excision of segments 22 to 39, 89 to 111, 119 to 133 and 191 to 197). In AMoRe, the 2MJP template structure is ranked in position 27<sup>th</sup> in the list, while the best ranking solutions corresponds to the full length models and after the pre-refinement process, 2 full length models comes first and are immediately followed by the 2MJP template structure (best solution  $R_{\text{free}}$ : 54.3,  $R_{\text{work}}$ : 44.7; 2MJP:  $R_{\text{free}}$ : 55.3  $R_{\text{work}}$ : 45.2). The amplitude of the R-factors and difference between that of the free and the working set prompt for a visual inspection of the corresponding electron density maps, based on for highest scoring model and on 2MJP itself, in order to validate the CaspR solution(s). It turned out that sole the solution based on the homology model provided a usable map for further refinement (most likely due to the presence of the correct side chains in the model).

### **BIGS protein YecD (1J2R) : an E. coli protein belonging to the isochorismatase family**

The *Escherichia coli* gene *yecD* encodes a 20 kD protein of unknown function annotated in Swiss-Prot as belonging to the isochorismatase family (Swiss-Prot: P37437). The crystals belong to the orthorhombic space group P2<sub>1</sub>2<sub>1</sub>2 with 4 molecules in the asymmetric unit. The YecD protein shares 19.5% and 24.6% identity with the 1IM5 and the 1NBA structures. Out of the 60 models generated and screened through the CaspR procedure, 30 correspond to truncated models (positions 1 to 18, 35 to 39 and 68 to 103). In the sorted list of the best ranking solutions in AMoRe, the 1NBA template is 13<sup>th</sup> and 1IM5 is 16<sup>th</sup>. The ten best solutions all correspond to truncated models and all converged through the pre-refinement process while 1NBA and 1IM5 did not (best solution:  $R_{\text{free}}$ : 50.6,  $R_{\text{work}}$ : 42.0; 1NBA:  $R_{\text{free}}$ : 58.0  $R_{\text{work}}$ : 56.5; 1IM5:  $R_{\text{free}}$ : 59.6  $R_{\text{work}}$ : 57.1). This case is actually the first structure that has been solved entirely using the method now automated in CaspR.

### **REFERENCES**

1. Backbro, K., Lowgren, S., Osterlund, K., Atepo, J., Unge, T., Hulten, J., Bonham, N. M., Schaal, W., Karlen, A. & Hallberg, A. (1997) Unexpected binding mode of a cyclic sulfamide HIV-1 protease inhibitor. *J. Med. Chem.*, **40**, 898-902.
2. Spinelli, S., Liu, Q. Z., Alzari, P. M., Hirel, P. H. & Poljak, R. J. (1991) The three-dimensional structure of the aspartyl protease from the HIV-1 isolate BRU. *Biochimie*, **73**, 1391-1396.
3. Korolev, S., Koroleva, O., Petterson, K., Gu, M., Collart, F., Dementieva, I. & Joachimiak, A. (2002) Autotracing of E. Coli Acetate Coa Transferase A-Subunit Structure Using 3.4 Å MAD and 1.9 Å Native Data. *Acta Cryst.*, **D58**, 2116-2121.

4. Sanghani, P. C., Robinson, H., Bennett-Lovsey, R., Hurley, T. D. & Bosron, W. F. (2003) Structure-Function Relationships in Human Class III Alcohol Dehydrogenase (Formaldehyde Dehydrogenase). *Chem. Biol. Interact.*, **143**, 195-200.