W152–W157 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl187

$(PS)^2$: protein structure prediction server

Chih-Chieh Chen[1], Jenn-Kang Hwang[1,2,3] and Jinn-Moon Yang[1,2,3,*]

[1]Institute of Bioinformatics, National Chiao Tung University, Hsinchu, 30050, Taiwan, [2]Department of Biological Science and Technology, National Chiao Tung University, Hsinchu, 30050, Taiwan and [3]Core Facility for Structural Bioinformatics, National Chiao Tung University, Hsinchu, 30050 Taiwan

Received February 14, 2006; Revised and Accepted March 8, 2006

ABSTRACT

Protein structure prediction provides valuable insights into function, and comparative modeling is one of the most reliable methods to predict 3D structures directly from amino acid sequences. However, critical problems arise during the selection of the correct templates and the alignment of query sequences therewith. We have developed an automatic protein structure prediction server, $(PS)^2$, which uses an effective consensus strategy both in template selection, which combines PSI-BLAST and IMPALA, and target–template alignment integrating PSI-BLAST, IMPALA and T-Coffee. $(PS)^2$ was evaluated for 47 comparative modeling targets in CASP6 (Critical Assessment of Techniques for Protein Structure Prediction). For the benchmark dataset, the predictive performance of $(PS)^2$, based on the mean GTD_TS score, was superior to 10 other automatic servers. Our method is based solely on the consensus sequence and thus is considerably faster than other methods that rely on the additional structural consensus of templates. Our results show that $(PS)^2$, coupled with suitable consensus strategies and a new similarity score, can significantly improve structure prediction. Our approach should be useful in structure prediction and modeling. The $(PS)^2$ is available through the website at http://ps2.life.nctu.edu.tw/.

INTRODUCTION

In the post-genomics era, one of the major challenges facing the structural biology research community is to determine the biological functions of genes identified through large-scale sequencing efforts. Knowledge of the 3D structure of a protein is crucial for understanding the molecular basis of its function. Unfortunately, the gap between the number of solved protein structures and the number of protein sequences continues to widen rapidly due to the long and expensive processes required for solving structures experimentally. Computational prediction of structures from amino acid sequence is an emerging and promising method that may help to narrow this gap. These methods have great potential to approximate the structure of newly acquired sequences based on known structures of similar sequence available from the rapidly growing number of protein crystal structures.

Comparative modeling generally comprises four main steps: (i) searching and selecting at least one known protein structure (the template) that is similar to the query (target sequence); (ii) alignment of the target sequence and the template(s); (iii) building models based on the chosen template(s); and (iv) evaluating the models. These steps can be reiterated until a satisfactory model structure is achieved. Currently, the first two steps are considered most critical because the accuracy of comparative models often tends to increase with the target–template sequence identity and the correctness of the alignment. A number of servers have been developed for automated comparative modeling (1–8). Several servers that yield predictions based on a set of different methods have demonstrated that consensus methods are significantly better than individual methods with regard to comparative modeling (3) and fold recognition (5). However, these methods have focused on target–template alignments and final model selections.

Here, we report the development of an automatic protein structure prediction server, $(PS)^2$, using a consensus strategy applied both in the template search/selection and target–template alignment phase. $(PS)^2$ was tested for all comparative modeling targets (47 targets) in CASP6 (Critical Assessment of Techniques for Protein Structure Prediction) (9). Our consensus procedure is computationally efficient and scalable to a greater number of combinations. Our experimental results demonstrate improved prediction accuracy relative to other automatic servers based on GTD_TS score (9,10).

METHODS AND IMPLEMENTATION

The efficiency of $(PS)^2$ derives from the ability to use an effective consensus strategy both in template selection [PSI-BLAST (11) and IMPALA (12)] and target–template alignment [PSI-BLAST, IMPALA and T-Coffee (13)]

*To whom correspondence should be addressed. Tel: +886 35712121-56942; Fax: +886 35729288; Email:moon@cc.nctu.edu.tw

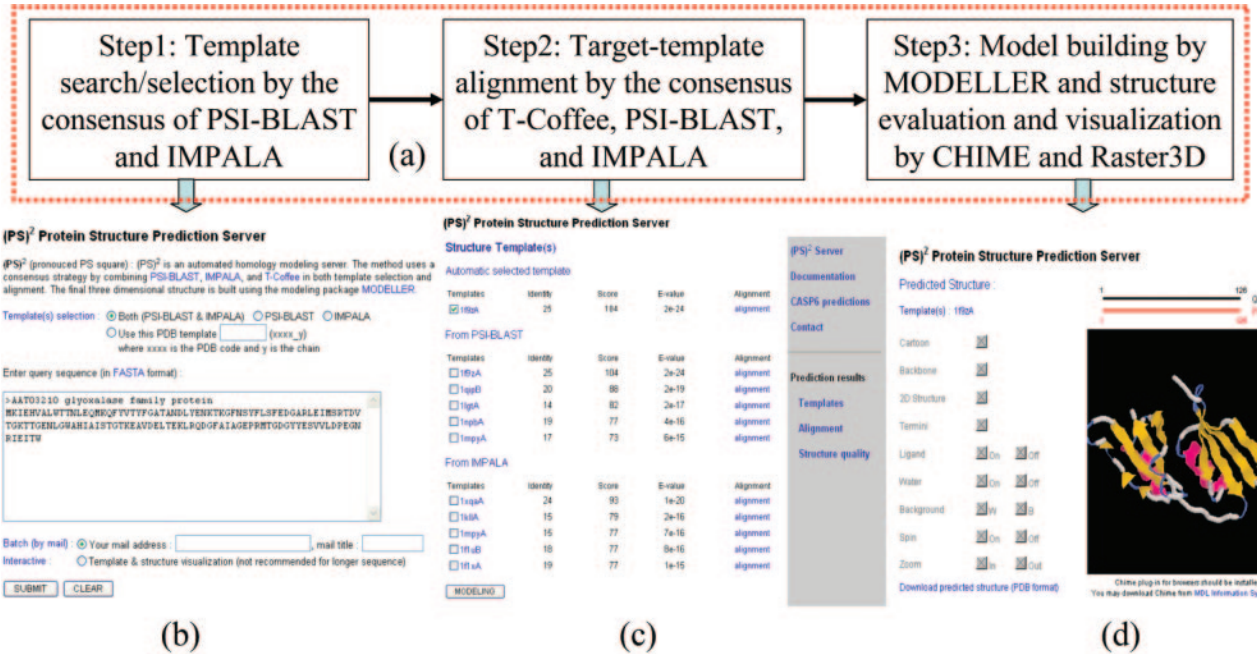

Figure 1. Overview of the $(PS)^2$ using the protein sequence of F2365 glyoxalase protein (AAT03210) in *L.monocytogenes* as query. (**a**) Main procedure; (**b**) The assignments of template selection method, target sequence, and interaction/batch; (**c**) Template selection (single/multiple templates) using the interaction module based on 20 candidates provided by PSI-BLAST and IMPALA; (**d**) Evaluation and visualization of the predicted structure.

(Figure 1). $(PS)^2$ comprises the following four steps: template selection, target–template alignment, model building, and model evaluation and visualization. These steps are repeated until a satisfactory model structure is achieved. The $(PS)^2$ consensus approach uses a set of publicly available tools for template search/selection and target–template alignment to produce the inputs for MODELLER (14), a comparative modeling tool based on the condition that spatial restraints must be satisfied.

For the easy use of the $(PS)^2$ server, it was designed to function with a minimum of user input, i.e., only the target sequence in FASTA format is needed (Figure 1b), and to provide 3D structure visualization directly through the web browser (Figure 1d). The server will automatically select suitable templates based on the consensus outputs of two profile search tools (e.g. PSI-BLAST and IMPALA). Alternatively, the user may specify a template structure. The automated modeling procedure begins when at least one modeling template is available. On the other hand, since comparative modeling procedures can have differential complexity, the $(PS)^2$ server provides for both interactive and batch modes (Figure 1b). In the interative mode, users can select different templates (Figure 1c) and visualize the modeled results (Figure 1d) on the $(PS)^2$ website until a satisfactory model structure is obtained. In the batch mode, $(PS)^2$ will automatically send the modeled results to users by Email when the automated modeling procedure is complete. The modeling procedure is briefly described in the following subsections.

Template search/selection

$(PS)^2$ uses the consensus of PSI-BLAST and IMPALA for the template search. PSI-BLAST and IMPALA are widely used for local sequence alignments with different profile search strategies. PSI-BLAST scans the profile of the query sequence against each of the template sequences in a database. In contrast, IMPALA searches the query sequence against each of the template profiles, which constitute a database of PSI-BLAST-generated position-specific score matrices (PSSMs). The template sequence library of $(PS)^2$ is extracted from the Protein Data Bank (PDB) (15). Any given pair of sequences in the library has <95% sequence identity. Currently, each template profile in the IMPALA profile library, which included 12 011 sequences, was constructed using PSI-BLAST by searching against the nrdb90 database.

For each protein sequence, $(PS)^2$ collected 20 templates from both the top 10 templates of PSI-BLAST and IMPALA by searching the template sequence and template profile library, respectively. $(PS)^2$ utilized a sequence similarity score (S_{IR}) which is a good template classifier if the optimal sequence alignment could be found. The sequence similarity score is given as $S_{IR} = (SI + AP)/2$ where SI is the sequence identity and AP is the alignment percentage between the query protein and the template sequence. As both PSI-BLAST and IMPALA are local alignment tools, the AP is important for selecting a right template. The AP is defined as the number of aligned residues divided by the total number of residues of a query protein sequence. Hence, from among the 20 templates $(PS)^2$ automatically selects the one with the highest S_{IR}, which is aligned by our consensus algorithm (Figure 2) using the resulting alignments of PSI-BLAST, IMPALA and T-Coffee (a multiple global sequence alignment tool).

Target–template alignment

As previous studies (1–8) indicate, the most persistent problem facing comparative modeling is probably the alignment of the query sequence with the template(s). With $(PS)^2$, we attempted

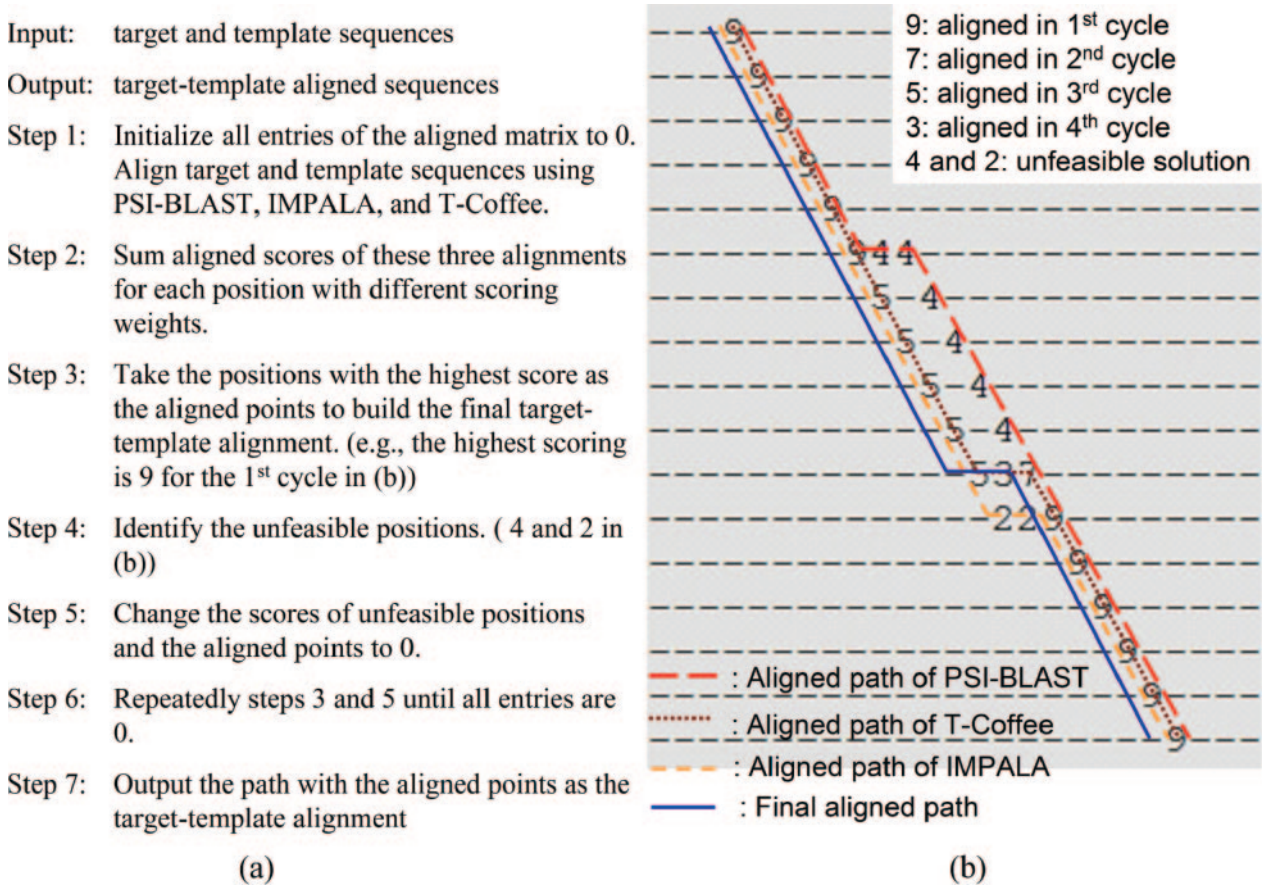

Figure 2. Template selection and target–template alignment in $(PS)^2$. (**a**) The consensus algorithm with T-Coffee, PSI-BLAST and IMPALA. (**b**) Example and unfeasible solutions.

to improve comparative modeling by considering alternative and consensus alignments based on the alignments of PSI-BLAST, IMPALA and T-Coffee. To efficiently combine the results of these alignment methods, we designed a consensus sequence algorithm (Figure 2) by considering the collective alignments from these tools and then given the target–template aligned-result. We briefly describe these steps as follows: (i) initialize all entries of the consensus matrix to 0; (ii) sum up aligned scores of these three alignments for each position with different scoring weights (IMPALA is 2, PSI-BLAST is 4 and T-Coffee is 3); (iii) take the positions with the highest score as the aligned points to build the final target–template alignment; (iv) identify the unfeasible positions (e.g. an amino acid in the target sequence is aligned with two amino acids in the template); (v) reset the scores of unfeasible positions and the aligned points in the consensus matrix to 0; (vi) repeat steps (iii–v) until all entries are 0; and (vii) output the path with the aligned points as the target–template alignment.

Model building and model evaluation

The final 3D protein structures were built from the consensus alignment using the homology-modeling package, MODELLER, which automatically calculated a model containing all non-hydrogen atoms using geometric restraints and molecular dynamic annealing. After MODELLER generated a predicted model with no other refinements, the program PROCHECK (16) was used to evaluate the quality of this model based on the *G*-factor. Finally, the predicted model was displayed by Raster3D (17) and automatically sent to users. The components of the $(PS)^2$ server shown in Figure 1 were built using PHP and Perl.

Input format

$(PS)^2$ is an easy-to-use web server (Figure 1b). Uses input the query protein sequence in FASTA format and choose a template selection method from 'Both', PSI-BLAST or IMPALA. The default option in template(s) selection is 'Both' which uses a consensus method combining PSI-BLAST and IMPALA. Users are also able to assign a specific PDB code as the template for the query sequence. Moreover, $(PS)^2$ provides both batch and interactive mode. In the batch mode, $(PS)^2$ automatically selects the template(s), while in the interative mode $(PS)^2$ allows the user to assign specific template(s) from a list of candidates (Figure 1c) Finally, the server sends the predicted results to the user's Email address.

Output format

Typically, the $(PS)^2$ server yielded a predicted structure within 5 min if the sequence length is ~200. The predicted results of the $(PS)^2$ server consists of the selected template(s), target–template alignment, predicted structure and structure evaluations (Figure 1d). The server provides the selected template and a list candidates yielded by PSI-BLAST and IMPALA (Figure 1c). The predicted structure is visualized in PNG format generated by MolScript (18) and Raster3D packages (17). If the Chime is installed in a browser, the output will display the predicted structure in the browser (Figure 1d). The server allows a user to download the predicted structure coordinates in the PDB format; furthermore, the target–template alignment in PIR format and the structure quality factors are also provided.

RESULTS

The global distance test_total score (GTD_TS) of Cα atoms was used to assess the correctness of the predicted model (10). GTD_TS has been commonly used in modeling studies and in the CASP community. GTD_TS is defined as

$$GDT\_TS = 100\frac{\sum_d GDT_d/N}{4}(\%) \quad d \in \{1,2,4,8\},$$

where N is the total number residues of a target, GDT_d is the number of aligned residues whose Cα-atom distance between the target and predicted model is less than d Å after superposition of the two structures; and d is 1, 2, 4 or 8 Å.

$(PS)^2$ was evaluated for 47 comparative modeling targets in CASP6 (9) (Figure 3) and the mean GDT-TS score was 66.69. In order to test $(PS)^2$ on these targets, each profile in the IMPALA profile library, which comprises 9775 sequences obtained from PDB on June 30, 2004, was constructed using PSI-BLAST by searching against the nrpb90 database. This server utilized the similarity score (S_{IR}) for template selection to improve prediction accuracy. When $(PS)^2$ used the template with the lowest expected value (E-value) in the hit structures similar to ESyPred3D (3), the mean GDT_TS scores for PSI-BLAST and IMPALA are 57.99 and 62.29, respectively. These scores were improved to 62.19 and 62.72, respectively, when the S_{IR} was applied in template selection. These experimental results show that S_{IR}, combining both the sequence identity and the alignment percentage, is a useful strategy for template selection since a low E-value does always imply a high sequent identity for cases such as T0229, T0231 and T0264. For target T0264, the sequence identity is 11.83% and GDT_TS score is 37.53 when the protein with the lowest PSI-BLAST E-value ($\sim 10^{-55}$) was used as the template (PDB code 1pjqA). In contrast, the sequence identity is 31.32%, the expected value is $\sim 10^{-39}$, and GDT_TS score is 64.97 when the protein with the highest S_{IR} was selected as the template (PDB code 1vhvA).

$(PS)^2$ outperformed PSI-BLAST and IMPALA alignments based on mean GDT_TS scores. PSI-BLAST and IMPALA selected the same templates for 32 targets among 47 targets and PSI-BLAST identified 10 better templates than IMPALA for 10 targets. Conversely, IMPALA identified five better templates than PSI-BLAST for five other targets. The experimental results show that the consensus sequence algorithm (Figure 2), combining both local and global multiple sequence alignment mechanisms, could indeed improve the performance. PSI-BLAST and IMPALA help to yield homologous protein sequences and local alignments by utilizing profile alignments, whereas T-Coffee expands local alignments to global alignments. For example, for T0205, the aligned percentages are 77.78% (PSI-BLAST), 79.80% (IMPALA), 100% (T-Coffee) and 100% (consensus method); moreover, the GDT_TS score are 66.94% (PSI-BLAST), 69.09% (IMPALA), 73.93% (T-Coffee) and 75.27% (consensus method).

Using these 47 targets, we compared the prediction accuracy of $(PS)^2$ with the 10 automatic servers (Figure 3). The mean GDT_TS scores of these 11 servers are 66.69 [$(PS)^2$], 64.92 [ROBETTA (2)], 63.14 [ESyPred3D (3)], 62.54 [3D-JIGSAW-recomb (1)], 61.27 [mGenTHREADER (7)], 61.08 [3D-JIGSAW-server (1)], 58.11 [PROSPECT (8)], 57.93 [Pmodeller5 (5)], 57.62 [PROTINFO (6)], 56.37

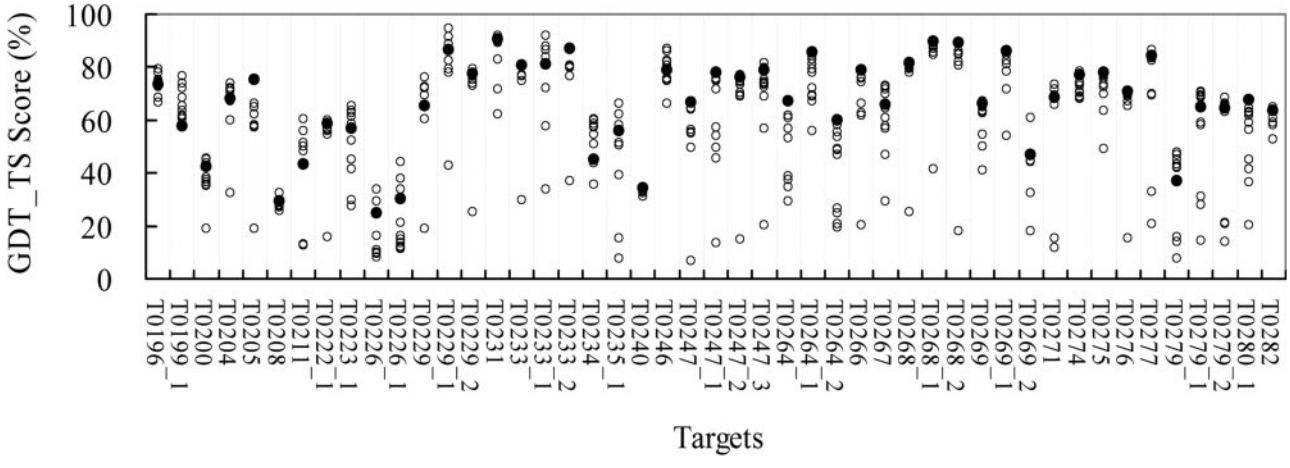

Figure 3. Comparison $(PS)^2$ (black) with 10 automatic servers of the prediction accuracies (GDT_TS scores) on 47 targets in CASP6. The results of these 10 automatic servers are summarized from http://predictioncenter.genomecenter.ucdavis.edu/casp6/Casp6.html.

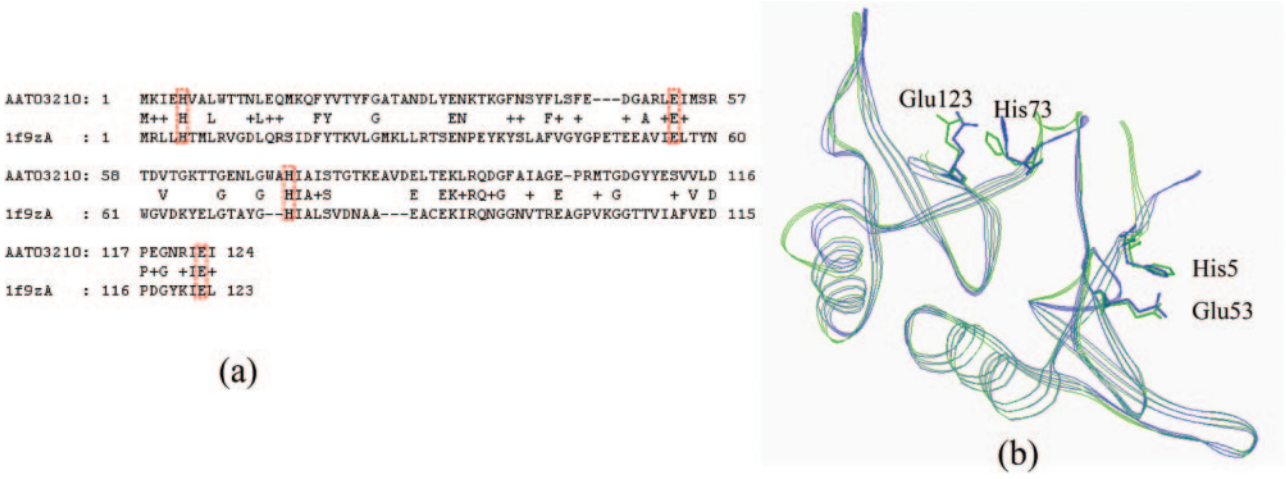

Figure 4. Predicted structure results of the $(PS)^2$ server using F2365 glyoxalase protein (AAT03210) sequence from *L.monocytogenes* as a query. (**a**) The alignment result between the query sequence and the selected template, glyoxalase I (PDB code 1f9z) from *E.coli*. (**b**) The structure alignment of the predicted structure (green) and the template (blue). Four important residues, which are responsible for the binding metal activity, are also shown.

[Pcons5 (5)] and 35.57 [Pcomb2 (5)]. For these targets, the mean GDT_TS score using $(PS)^2$ was superior to these of the 10 automatic servers; moreover, the individual GDT_TS scores from $(PS)^2$ were comparable. Using mean GDT_TS scores, $(PS)^2$ obtained 13 predicted structures in the first rank and 8 structures in the second place. These analysis results suggest that the accuracy of $(PS)^2$ is comparable with those of previous prediction servers.

Example analysis

$(PS)^2$ predicted the structure of the F2365 glyoxalase protein (AAT03210) sequence in *Listeria monocytogenes* (19) (Figure 4). It selected the native structure of glyoxalase I (GlxI) (PDB code 1f9zA) from *Escherichia coli* (20) as the template. GlxI is the first of two enzymes in the pathway to converts cytotoxic α-keto aldehydes into nontoxic α-hydroxy-carboxylic acids. This pathway is important in that an increase in methylglyoxal can produce toxic effects by reacting with DNA, RNA, and proteins. Therefore, GlxI has been utilized in the design of anticancer and antimalarial agents (21).

The template shares 23.7% sequence identity with the query sequence and the target–template alignment is shown in Figure 4a. $(PS)^2$ automatically aligned four important residues together (His5, Glu53, His73 and Glu123 in the query sequence; His5, Glu56, His74 and Glu122 in the template sequence), which are responsible for the binding metal activity of the GlxI family (red blocks in Figure 4a). The superimposing result (Figure 4b) of the predicted structure (green) and the template structure (blue) also shows that the coordinates of side chains and backbones of these four residues are similar.

This protein sequence (AAT03210) was also submitted to SWISS-MODEL (22), which is a widely used homology-modeling server and ESyPred3D. SWISS-MODEL is unable to find a suitable template since no sequences above 25% sequence identity are found. On the other hand, ESyPred3D selected a native structure of GlxI [PDB code 1qipD (23)] from *Homo sapiens* as template. The template shares 17.4% identity with this query sequence. Human GlxI is active in the presence of Zn^{2+}; but *E.coli* GlxI is inactive in the presence of Zn^{2+} and is maximally active with Ni^{2+}, as *L.monocytogenes* GlxI (AAT03210) does. (19). These analysis results show that the query sequence is more correlated to *E.coli* GlxI than Human GlxI.

CONCLUSION

The key novelty of $(PS)^2$ is the seamless ability of blending local and global multiple sequence alignment mechanisms to allow them to work cooperatively by a new similarity score (S_{IR}). The analysis using $(PS)^2$ was significantly faster because $(PS)^2$ uses an effective consensus strategy that combines three publicly available tools installed on the same machine; moreover, $(PS)^2$ is based solely on the consensus sequence and thus is considerably faster than other methods that rely on the additional structural consensus of templates. We believe that $(PS)^2$ is a fast homology-modeling server and should be useful in structure prediction and modeling.

ACKNOWLEDGEMENT

We are grateful to both the hardware and software supports of the Structural Bioinformatics Core Facility at National Chiao Tung University. J.-M. Yang was supported by National Science Council and the University System at Taiwan-Veteran General Hospital Grant. J.-K. Hwang was supported by National Science Council, National Research Program of Genomic Medicine, and the University System at Taiwan-Veteran General Hospital Grant. Funding to pay the Open Access publication charges for this article was provided by National Science Council.

Conflict of interest statement. None declared.

REFERENCES

1. Bates,P.A., Kelley,L.A., MacCallum,R.M. and Sternberg,M.J.E. (2001) Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. *Proteins*, **Suppl 5**, 39–46.
2. Chivian,D., Kim,D.E., Malmstrom,L., Bradley,P., Robertson,T., Murphy,P., Strauss,C.E.M., Bonneau,R., Rohl,C.A. and Baker,D. (2003) Automated prediction of CASP-5 structures using the robetta server. *Proteins*, **53**, 524–533.
3. Lambert,C., Leonard,N., Bolle,X.D. and Depiereux,E. (2002) EsyPred3D: prediction of proteins 3D structures. *Bioinformatics*, **18**, 1250–1256.
4. Ogata,K. and Umeyama,H. (2000) An automatic homology modeling method consisting of database searches and simulated annealing. *J. Mol. Graph. Model.*, **18**, 258–272.
5. Wallner,B., Fang,H. and Elofsson,A. (2003) Automatic consensus-based fold recognition using Pcons, ProQ, and Pmodeller. *Proteins*, **53**, 534–541.
6. Hung,L.-H. and Samudrala,R. (2003) PROTINFO: secondary and tertiary protein structure prediction. *Nucleic Acids Res.*, **31**, 3296–3299.
7. Jones,D.T. (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.*, **287**, 797–815.
8. Xu,Y. and Xu,D. (2000) Protein threading using PROSPECT: design and evaluation. *Proteins*, **40**, 343–354.
9. Tress,M., Ezkurdia,I., Grana,O., Lopez,G. and Valencia,A. (2005) Assessment of predictions submitted for the CASP6 comparative modeling category. *Proteins*, **S7**, 27–45.
10. Zemla,A. (2003) LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res.*, **31**, 3370–3374.
11. 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.
12. Schaffer,A.A., Wolf,Y.I., Ponting,C.P., Koonin,E.V., Aravind,L. and Altschul,S.F. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics*, **15**, 1000–1011.
13. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.
14. Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 799–815.
15. Deshpande,N., Addess,K.J., Bluhm,W.F., Merino-Ott,J.C., Townsend-Merino,W., Zhang,Q., Knezevich,C., Xie,L., Chen,L., Feng,Z. *et al.* (2005) The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res.*, **33**, D233–D237.
16. Laskowski,R.A., MacArthur,M.W., Moss,D.S. and Thornton,J.M. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.*, **26**, 283–291.
17. Merritt,E.A. and Bacon,D.J. (1997) Raster3D: photorealistic molecular graphics. *Methods Enzymol.*, **277**, 505–524.
18. Kraulis,P.J. (1991) MOLSCRIPT:a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.*, **24**, 946–950.
19. Nelson,K.E., Fouts,D.E., Mongodin,E.F., Ravel,J., DeBoy,R.T., Kolonay,J.F., Rasko,D.A., Angiuoli,S.V., Gill,S.R., Paulsen,I.T. *et al.* (2004) Whole genome comparisons of serotype 4b and 1/2a strains of the food-borne pathogen *Listeria monocytogenes* reveal new insights into the core genome components of this species. *Nucleic Acids Res.*, **32**, 2386–2395.
20. He,M.M., Clugston,S.L., Honek,J.F. and Matthews,B.W. (2000) Determination of the structure of *Escherichia coli* Glyoxalase I suggests a structural basis for differential metal activation. *Biochemistry*, **39**, 8719–8727.
21. Kavarana,M.J., Kovaleva,E.G., Creighton,D.J., Wollman,M.B. and Eiseman,J.L. (1999) Mechanism-based competitive inhibitors of glyoxalase i: intracellular delivery, in vitro antitumor activities, and stabilities in human serum and mouse serum. *J. Med. Chem.*, **42**, 221–228.
22. Schwede,T., Kopp,J., Guex,N. and Peitsch,M.C. (2003) SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res.*, **31**, 3381–3385.
23. Cameron,A.D., Ridderstrom,M., Olin,B., Kavarana,M.J., Creighton,D.J. and Mannervik,B. (1999) Reaction mechanism of glyoxalase I explored by an X-ray crystallographic analysis of the human enzyme in complex with a transition state analogue. *Biochemistry*, **38**, 13480–13490.