Journal of
**Applied
Crystallography**

ISSN 0021-8898

The PRIDE server for protein three-dimensional similarity

Kristian Vlahovicek, Oliviero Carugo and Sándor Pongor

Journal of
**Applied
Crystallography**

ISSN 0021-8898

Received 8 July 2002
Accepted 17 July 2002

The PRIDE server for protein three-dimensional similarity

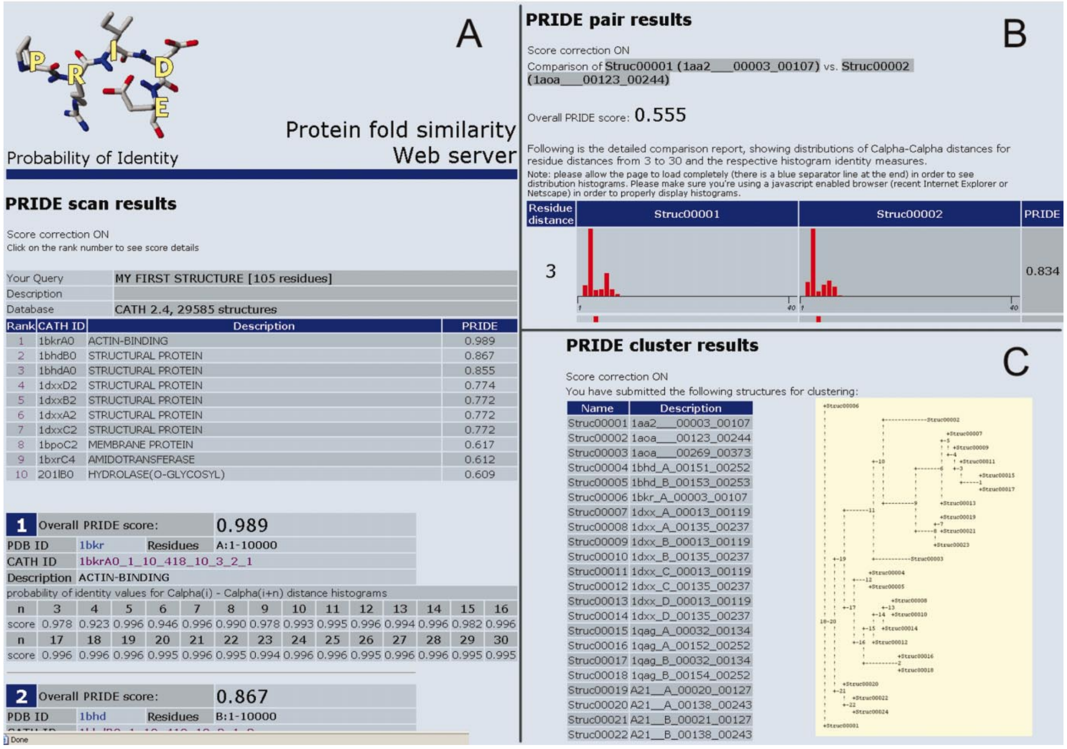
Kristian Vlahovicek,[a] Oliviero Carugo[a,b]* and Sándor Pongor[a]

[a]Protein Structure and Function Group, International Centre for Genetic Engineering and Biotechnology, Area Science Park, Padriciano 99, 34012 Trieste, Italy, and [b]Department of General Chemistry, University of Pavia, Viale Taramelli 12, 27100 Pavia, Italy. Correspondence e-mail: carugo@icgeb.trieste.it

The PRIDE server is an implementation of the *PRIDE* algorithm that compares protein three-dimensional structures in terms of their C^{α} distance distributions. In response to queries presented as single or concatenated Protein Data Bank (PDB) files, the server can carry out (i) a pairwise comparison of two protein three-dimensional structures, (ii) a structural clustering of protein three-dimensional structures, providing a distance matrix and a dendrogram as an output; and (iii) a similarity search with a protein domain structure query against the CATH database.

1. Introduction

The recognition of the similarity between protein three-dimensional structures is crucially important in molecular evolution studies, in function prediction methods as well as in the quality assessment of three-dimensional structure prediction. Traditionally, the similarity between a pair of protein three-dimensional structures is evaluated either by the alignment of the distance matrices or by structural superposition (for a review see Lesk, 2002), though methods based on various other simplified structural representations have also been developed (Johnson & Lehtonen, 2000). Among the latter, a novel procedure has recently been proposed (Carugo & Pongor, 2002) in which a protein structure is represented by a set of 28 histograms, each describing the distribution of the $C^{\alpha}(i)$–$C^{\alpha}(i+n)$ distances ($3 \leq n \leq 30$). In order to estimate the similarity of two protein structures, two sets of 28 histograms are compared in a pairwise manner *via* chi-square contingency table analysis (Dowdy & Wearden, 1991), and the resulting 28 probability values (PI) are averaged to give an overall probability of identity (PRIDE) score ($0 \leq \text{PRIDE} \leq 1$). In comparison with other methods of structural comparison, it is worth mentioning that the computation of the PRIDE score is extremely



Probability of Identity

Protein fold similarity
Web server

PRIDE scan results

Score correction ON
Click on the rank number to see score details

Your Query	MY FIRST STRUCTURE [105 residues]
Description	
Database	CATH 2.4, 29585 structures

Rank	CATH ID	Description	PRIDE
1	1bkrA0	ACTIN-BINDING	0.989
2	1bhdB0	STRUCTURAL PROTEIN	0.867
3	1bhdA0	STRUCTURAL PROTEIN	0.855
4	1dxxD2	STRUCTURAL PROTEIN	0.774
5	1dxxB2	STRUCTURAL PROTEIN	0.772
6	1dxxA2	STRUCTURAL PROTEIN	0.772
7	1dxxC2	STRUCTURAL PROTEIN	0.772
8	1bpoC2	MEMBRANE PROTEIN	0.617
9	1bxrC4	AMIDOTRANSFERASE	0.612
10	201lB0	HYDROLASE(O-GLYCOSYL)	0.609

1 Overall PRIDE score:		0.989	
PDB ID	1bkr	Residues	A:1-10000
CATH ID	1bkrA0_1_10_418_10_3_2_1		
Description	ACTIN-BINDING		

probability of identity values for Calpha(i) - Calpha(i+n) distance histograms

n	3	4	5	6	7	8	9	10	11	12	13	14	15	16
score	0.978	0.923	0.996	0.946	0.996	0.990	0.978	0.993	0.995	0.996	0.994	0.996	0.982	0.996
n	17	18	19	20	21	22	23	24	25	26	27	28	29	30
score	0.996	0.996	0.996	0.995	0.996	0.995	0.994	0.996	0.996	0.995	0.996	0.996	0.995	0.995

2 Overall PRIDE score:		0.867	
PDB ID	1bhd	Residues	B:1-10000

Done

PRIDE pair results

Score correction ON
Comparison of Struc00001 (1aa2___00003_00107) vs. Struc00002 (1aoa___00123_00244)

Overall PRIDE score: 0.555

Following is the detailed comparison report, showing distributions of Calpha-Calpha distances for residue distances from 3 to 30 and the respective histogram identity measures.

Note: please allow the page to load completely (there is a blue separator line at the end) in order to see distribution histograms. Please make sure you're using a javascript enabled browser (recent Internet Explorer or Netscape) in order to properly display histograms.

Residue distance	Struc00001	Struc00002	PRIDE
3			0.834

PRIDE cluster results

Score correction ON
You have submitted the following structures for clustering:

Name	Description
Struc00001	1aa2___00003_00107
Struc00002	1aoa___00123_00244
Struc00003	1aoa___00269_00373
Struc00004	1bhd_A_00151_00252
Struc00005	1bhd_B_00153_00253
Struc00006	1bkr_A_00003_00107
Struc00007	1dxx_A_00013_00119
Struc00008	1dxx_A_00135_00237
Struc00009	1dxx_B_00013_00119
Struc00010	1dxx_B_00135_00237
Struc00011	1dxx_C_00013_00119
Struc00012	1dxx_C_00135_00237
Struc00013	1dxx_D_00013_00119
Struc00014	1dxx_D_00135_00237
Struc00015	1qag_A_00032_00134
Struc00016	1qag_A_00152_00252
Struc00017	1qag_B_00032_00134
Struc00018	1qag_B_00154_00252
Struc00019	A21__A_00020_00127
Struc00020	A21__A_00138_00243
Struc00021	A21__B_00021_00127
Struc00022	A21__B_00138_00243

Figure 1
The output of PRIDE_scan (*A*), PRIDE_pair (*B*) and PRIDE_cluster (*C*).

fast, and in addition, PRIDE is a metric in the mathematical sense and can thus be used in cluster analyses or other classification tasks. Structural classification based on PRIDE values is in very good agreement with the CATH classification scheme (Carugo & Pongor, 2002).

2. Implementation

The input of the PRIDE server is a protein structure in PDB format, which must begin with the HEADER field (which is used for identification) and finish with END. The ATOM lines of the C^α atoms must be present in the input; all other lines are ignored by the program. In some cases (see below) concatenated PDB files are used; in this case the HEADER-to-END chunks must be concatenated into a unique file. The input file is up-loaded directly form the user's computer. The server offers three main options.

(i) The PRIDE_pair option allows the comparison of two structures, submitted either independently or as a concatenated PDB file. The output includes the final PRIDE value as well a graphical rendering of the 28 distance histogram pairs along with the individual probability values resulting form their pairwise comparison, which allows a critical examination of the results. For example, a relatively high PRIDE value that results from only a few high values obtained between sequentially close C^α atoms can be a result of similar secondary structure content, rather than an indicator of a similar fold.

(ii) The PRIDE_cluster option is an all-against-all comparison of structures submitted in a concatenated PDB file. Three types of results are produced: (*a*) a distance matrix where each element $x(i, j)$, equals (1-PRIDE), is calculated between the *i*th and the *j*th structure submitted; (*b*) an ASCII dendrogram produced by nearest-neighbour clustering using the *NEIGHBOR* program (Felsenstein, 1995); (iii) a so-called Newick Standard Format tree file that can be used to re-plot the dendrogram for publication purposes using programs like *NJPLOT* (http://acnuc.univ-lyon1.fr/phylogeny/njplot) or *TREEVIEW* (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html), *etc.*

(iii) The PRIDE_scan option is a similarity search against the protein folds of release 2.4 of the CATH database (Orengo *et al.*, 1997). The result is a list of the database entries ranked according to their PRIDE similarity to the query (Fig. 1). In the output list, the entry names are linked to the respective CATH and PDB records. There is an option to view the 28 individual PI probability values that underlie the PRIDE value used for ranking. Domain collections other than the CATH database will be implemented in the future. It is noted that the query used to scan a domain collection must be a domain structure, and not a structure consisting of several domains. Multidomain proteins need to be divided into their constituent domains and analysed separately.

In all three cases there is an option to correct the bias resulting from accidental similarities that may randomly occur between proteins of different size. This correction is based on empirical observation, and was found to improve the performance slightly, especially in the case of database scanning. The options are described in a series of help files.

The main advantage of using PRIDE to characterize the similarity between protein structures is the possibility of automatic structural clustering and fast database search. At present, comparing a query of 100 amino acids with the 30000 entries of the CATH database (Orengo *et al.*, 1997) takes about 9 s using a PC equipped with a 1.3 GHz AMD Athlon CPU, thus permitting the service to be fully interactive.

3. Availability

The server may be accessed free of charge by anyone at the URL http://www.icgeb.org/pride. A detailed on-line manual (PRIDE help) is also available.

This work was partly funded by the EU project ORIEL (IST-2001-32688) 'On-Line Research Environment for the Life Sciences', coordinated by the European Molecular Biology Organization (EMBO).

References

Carugo, O. & Pongor, S. (2002). *J. Mol. Biol.* **315**, 887–898.
Dowdy, S. & Wearden, S. (1991). *Statistics for Research*. New York: Wiley.
Felsenstein, J. (1995). *PHYLIP* (*Phylogeny Inference Package*), 3,75c edit. Department of Genetics, University of Washington, Seattle, USA.
Johnson, M. S. & Lehtonen, J. V. (2000). *Bioinformatics: Sequence, Structure and Databanks*, edited by D. Higgins & W. Taylor, pp. 15–50. Oxford University Press.
Lesk, A. M. (2002). *Introduction to Bioinformatics*. Oxford University Press.
Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). *Structure*, **5**, 1093–1108.