BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 15 2005, pages 3322–3323
doi:10.1093/bioinformatics/bti513

Structural bioinformatics

Efficient recognition of folds in protein 3D structures by the improved PRIDE algorithm

Zoltán Gáspári[1,2], Kristian Vlahovicek[3] and Sándor Pongor[1,3,*]

[1]Bioinformatics Group, Biological Research Center, Hungarian Academy of Sciences, Temesvári krt. 62, Szeged, Hungary, [2]Department of Organic Chemistry, Eötvös Loránd University, Pázmány Péter sétány 1/A, Budapest, Hungary and [3]Protein Structure and Bioinformatics, International Centre for Genetic Engineering and Biotechnology, Area Science Park, Padriciano 99, 34012 Trieste, Italy

Received on March 23, 2005; revised on May 18, 2005; accepted on May 19, 2005
Advance Access publication May 24, 2005

ABSTRACT

Summary: An improved version of the PRIDE (PRobaility of IDEntity) fold prediction algorithm has been developed, based on more solid statistical basis, fast search capabilities and efficient input structure processing. The new algorithm is effective in identifying protein structures at the 'H' level of the CATH hierarchy.

Availability: The new algorithm is integrated into the PRIDE2 web servers at http://pride.szbk.u-szeged.hu and http://www.icgeb.org/pride

Contact: pongor@icgeb.org

Supplementary information: Detailed documentation and performance evaluation is available in the description section of the PRIDE2 web server.

1 INTRODUCTION

The NP-hardness of the protein structure comparison problem inspired a number of structure comparison methods. More rigorous methods use structural alignment; fast methods are usually based on specific structural descriptions designed for quick comparison (for a recent review see Sierk and Kleywegt, 2004). The PRIDE (Probability of IDEntity) algorithm (Carugo and Pongor, 2002; Vlahovicek *et al*., 2002) falls into this second category. It is based on representing protein structures in terms of $C\alpha_i$–$C\alpha_{i+n}$ $(2 < n \leq 30)$ distance distributions, and comparing two sets of distributions (representing two protein structures, respectively) via contingency table analysis. Fold identification by PRIDE is based on nearest-neighbour analysis using the CATH database (Orengo *et al*., 1997). Even though the method is quite fast and the initial accuracy estimates were encouraging (Carugo and Pongor, 2002), PRIDE did not fare well in a recent evaluation of fold-identification servers, especially when compared with rigorous methods based on structural alignment (Novotny *et al*., 2003).

In this paper we describe a number of simple improvements to the original PRIDE algorithm that allowed us to significantly increase the prediction power of the method, without sacrificing speed.

2 RESULTS AND DISCUSSION

The changes implemented were designed to serve three general purposes: (1) increasing the accuracy, (2) increasing the speed and (3) simplifying the use of the server.

- The comparison of distributions is now carried out with the Kuiper variant of the Kolmogorov–Smirnov (KS) test (Press *et al*., 1992) which is a more robust— and in our case—a more sensitive method than the comparison of binned histograms using contingency table analysis.
- A fast two-step fold-identification method has been implemented in which the query is first compared with cumulative distributions of CATH topology groups. The 10 best groups are retained and then the structure is compared to the representatives of these groups only.
- A part of the mispredictions was found to be related to the fact that PRIDE does not identify substructure similarities such as partial structural alignments. A configurable window-sliding option (similar to the approach used by Gáspári *et al*., 2004) has been employed that provides a partial solution to this problem.
- Improved Protein Data Bank (PDB) file processing facilities are now implemented that can handle files with multiple chains, concatenated PDB files as well as files with missing coordinates.
- Local domain similarities are presented in a graphical form.
- Fold identification is based on a subset of the CATH version 2.5.1 database (Orengo *et al*., 1997). This has been constructed by retaining only one (if possible, the longest) structure at the 7th level of the CATH hierarchy, yielding a total of 17 844 structures. The group distributions named above were constructed by pooling distributions at the same 'H' level.
- It is now possible to search a subset of PDB database (Berman *et al*., 2000) which is derived from the 25% similarity-filtered list of the 2004 October release of PDB SELECT (Hobohm *et al*., 1992; Hobohm and Sander, 1994) yielding a total of 2485 structures.

The server program was written in PERL and C++. The server has three main options: (1) Simple structure comparison and clustering. These are now based on the KS test. (2) Fold identification using a subset of the CATH database. This is carried out either by a two-step

[*]To whom correspondence should be addressed.

Table 1. Correct predictions based on the benchmark dataset of Novotny *et al*. (2003)

Type	No. of structures	Original PRIDE[a]	PRIDE2 default[b]	PRIDE2 user-optimized[c]
All α	19	14	12	17
All β	19	14	18	18
$\alpha + \beta$	15	7	13	13
Few SS	8	3	8	8
Total (%)	61 (100)	38 (62)	51 (84)	56 (92)
CPU time (s)		1–2	10–12[d]	10–600

[a] Data taken from Novotny *et al*. (2003).
[b] Using default window/slide parameters of 160/80 and two-step search.
[c] Separately parametrized for each query.
[d] Approximately 1–3 s per individual query (substructure).

method, or a more thorough direct comparison with the database. For a typical query of 160 amino acids, the estimated CPU time (on a 900 MHz AMD Athlon machine) ranges from 3 s (two-step procedure) to 30 s (one-step procedure). (3) Comparison to PDBselect can also be carried out, the CPU time being ~3 s per query. Even though the speed is somewhat slower than that of the original PRIDE algorithm, the analysis is fast enough for on-line use and can be implemented on a single Linux-based PC. Detailed on-line help files have been added to the server.

The accuracy of fold prediction was tested on the set of structures used by Novotny *et al*. (2003) in their comparison of protein fold similarity servers. This set included 61 PDB structures that contained examples of CATH domains falling into the four major structural classes (mainly alpha, mainly beta, alpha + beta, a few secondary structures) (Novotny *et al*., 2003). The results summarized in Table 1 show a substantial improvement when compared with the previous version of PRIDE. According to the data of Novotny and co-workers, DALI (Holm and Sander, 1993) and CE (Shindyalov and Bourne, 1998) reached a success rate of 90 and 93%, respectively, on the same test set which compares quite well with the 84% result of PRIDE2, especially if the run times are also considered. The performance of PRIDE2 can be improved to 92% if the user selects individual window/slide parameters, or in some cases, uses full database search (Table 1). In particular, PRIDE performs well if the user submits fragments of a larger protein, rather than the protein itself. A detailed evaluation of the tests—including receiver operating characteristic curves—is available in the evaluation section at the PRIDE website.

Summarizing, the performance of PRIDE falls somewhat short of that of structural alignment algorithms, and this is in our opinion owing to the fact that PRIDE misses some of the all-alpha structures, especially if they are part of larger proteins. At the moment, PRIDE is more suited for interactive use; and it gives the best results if the approximate domain boundaries/sizes are a priori known. We hope that the speed of the analysis will make PRIDE competitive in large-scale applications.

ACKNOWLEDGEMENTS

The kind help of Vilmos Ágoston, Sergiu Netotea and Zoltán Hegedüs in installing the server is gratefully acknowledged. The work was supported by the Hungarian Office of Research and Development (OMFB-01887/2002, OMFB-00299/2002). S.P. is recipient of the Szent-Györgyi Award for teaching at the Department of Genetics and Molecular Biology, University of Szeged.

Conflict of Interest: none declared.

REFERENCES

Berman,H.M. *et al*. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Carugo,O. and Pongor,S. (2002) Protein fold similarity estimated by a probabilistic approach based on Cα–Cα distance comparison. *J. Mol. Biol.*, **315**, 887–898.

Gáspári,Z. *et al*. (2004) A simple fold with variations: the pacifastin inhibitor family. *Bioinformatics*, **20**, 448–451.

Hobohm,U. and Sander,C. (1994) Enlarged representative set of protein structures. *Protein Sci.*, **3**, 522–524.

Hobohm,U. *et al*. (1992) Selection of representative protein data sets from the Protein Data Bank. *Protein Sci.*, **1**, 409–417.

Holm,L. and Sander,C. (1993) Protein structure comparison by the alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.

Novotny,M. *et al*. (2003) Evaluation of protein fold comparison servers. *Proteins*, **54**, 260–270.

Orengo,C.A. *et al*. (1997) CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.

Press,W.H., Teukolsky,S.A., Vetterling,W.T and Flannery,B.P. (1992) *Numerical Recipes in C*, 2nd edn. Cambridge University Press, Cambridge.

Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.

Sierk,M.L. and Kleywegt,G.L. (2004) *Déjà vu* all over again: finding and analyzing protein structure similarities. *Structure*, **12**, 2103–2111.

Vlahovicek,K. *et al*. (2002) The PRIDE server for ptotein three-dimensional similarity. *J. Appl. Crystallogr.*, **35**, 648–649.