BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 2 2008, pages 272–275
doi:10.1093/bioinformatics/btm564

Structural bioinformatics

Fast protein fold estimation from NMR-derived distance restraints

Annamária F. Ángyán[1], András Perczel[1,2], Sándor Pongor[3,4] and Zoltán Gáspári[1,*]

[1]Institute of Chemistry, Eötvös Loránd University, [2]MHAS-ELTE Protein Modelling Group, Pázmány Péter sétány 1/A, 1117 Budapest, Hungary, [3]Protein Structure and Bioinformatics Group, International Centre for Genetic Engineering and Biotechnology, Padriciano 99, 34012 Trieste, Italy and [4]Bioinformatics Group, Biological Research Centre, Hungarian Academy of Sciences, Temesvári körút 62, 6701 Szeged, Hungary

Received on August 23, 2007; accepted on November 6, 2007

Advance Access publication November 13, 2007

Associate Editor: Burkhard Rost

ABSTRACT

Summary: PRIDE-NMR is a fast novel method to relate known protein folds to NMR distance restraints. It can be used to obtain a first guess about a structure being determined, as well as to estimate the completeness or verify the correctness of NOE data.

Availability: The PRIDE-NMR server is available at http://www.icgeb.org/pride

Contact: szpari@chem.elte.hu

Supplementary information: Description of the server and details of the tests presented can be found at http://www.icgeb.org/pride

1 INTRODUCTION

The main bottleneck in protein structure determination with NMR spectroscopy is structure calculation by using–primarily NOE-based–structural restraints derived from the acquired spectra. The length and outcome of this multi-step process depends heavily on the quality and quantity of the spectral data and also on the reliability of the resonance assignment. Although there are many approaches to speed up structure calculation, capable of yielding a protein structural model of acceptable quality (Herrmann *et al*., 2002; Rieping *et al*., 2007), thorough investigation of a chosen protein may require manual intervention by the researcher in order to separate valid experimental information from artifacts. Moreover, automated or semi-automated methods work best with high quality spectra not accessible for all proteins and conditions of interest. Information about secondary and tertiary structure can be obtained by analyzing chemical shifts (Cavalli *et al*., 2007) and residual dipolar couplings if a homolog with known 3D coordinates is available (Annila *et al*., 1999; Delaglio *et al*., 2000). However, as high-quality NMR structure determination relies primarily on NOE-based restraints, a fast method capable to relate known folds to the obtained NOE data set could be of valuable help for the NMR spectroscopist. Furthermore, even when structure determination is straightforward, an independent test of the validity of the obtained fold could be desirable.

Here we report the development of a conceptually simple and fast method, PRIDE-NMR, able to select folds compatible with a given set of NOE data. The name and concept comes from the fast protein fold comparison procedure, PRIDE (PRobability of IDEntity, Carugo and Pongor, 2002), which is based on the comparison of Cα–Cα distance distributions.

2 METHODS

PRIDE-NMR compares the distributions of short interproton distances (obtained from NMR experiments or back-calculated from 3D coordinates) within the widely defined protein backbone (amide H, Hα and Hβ atoms). The number of distance restraints or close H–H pairs is represented as a histogram with bins corresponding to the sequential separation of the participating residues (Fig. 1). The two histograms are compared with contingency analysis. Histograms are normalized to 100% and bins containing <5% of the total data are combined successively with the next ones to ensure than no values below 5% are used (described in detail in Carugo and Pongor, 2002). As in PRIDE, the resulting score ($0 \leq$ PRIDE-NMR Score ≤ 1) can be interpreted as the probability of the two data sets representing the same fold. The exclusion of side-chain hydrogen atoms beyond the β position renders the method largely independent of the sequences of the proteins compared.

Two approaches were introduced in order to increase the sensitivity: the first one is a score weighting with a given power (1, 2 or 3) of the ratio of the lengths of the proteins compared:

$$\text{PRIDE-NMR\_}W_x = \text{PRIDE-NMR} * \left(\frac{\textit{length of the shorter protein}}{\textit{length of the larger protein}}\right)^x$$

The second one is a filter that discards hits differing in length from the query by more than a chosen percentage. The size of the target protein (number of residues) is a piece of sequence-independent information known to the NMR-spectroscopist.

To set up a web server capable of relating NMR distance data sets to a wide range of known folds, we used the SCOP database (Murzin *et al*., 1995), which contains X-ray structures and also proteins with less than 40 residues, typically accessible to NMR structure determination. We used the 95% sequence similarity filtered subset of SCOP (also used in the Protein Classification Benchmark Database, Sonego *et al*., 2007). Hydrogen atoms were placed on all structures using the pdb2gmx program of the GROMACS molecular dynamics package (van der Spoel

*To whom correspondence should be addressed.

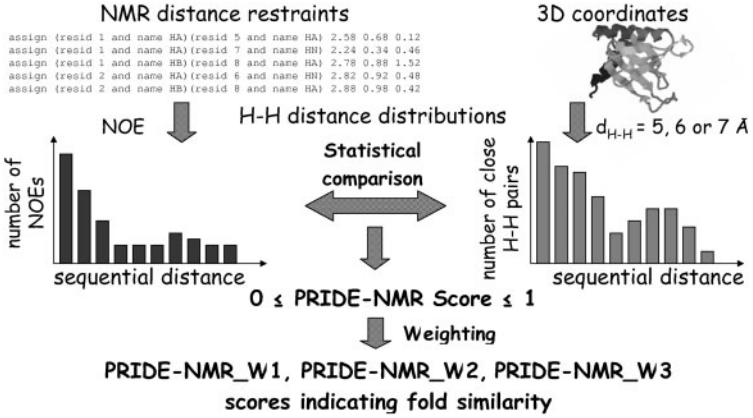

Fig. 1. Graphical representation of the PRIDE-NMR method.

et al., 2001) with the OPLS-AA force field (Kaminski *et al*., 2001). Minor modifications were made to enable handling of structures with missing atomic coordinates. H–H distance distributions were calculated with distance cutoffs of 5, 6 and 7 Å and averaging the positions of protons in alanine methyl groups. The server uses a database of precalculated distance distributions (back-calculated from 3D structures) and accepts distance restraints in X-PLOR/CNS format (Brünger *et al*., 1998). The user is allowed to choose the weighting and/or length filtration mode as well as the cutoff distance(s) for the database distributions (for multiple distances, the averaged scores are calculated). The server was implemented in Perl and C++ and is integrated into the PRIDE2 interface at http://www.icgeb.org/pride (Gáspári *et al*., 2005).

3 RESULTS

Owing to the relatively low computational demands of the implementation, the PRIDE-NMR server is extremely fast, yielding results in the order of a second. This speed allows for multiple runs with adjusting the parameters of the query to explore the relationship of an NOE data set to the folds in the SCOP database.

As a first test, five members from each of three protein families represented in SCOP with deposited NMR distance restraint sets were used: the ubiquitin-related (SCOP d.15.1.1; Table 1), the SH3-domain (SCOP b.34.2.1) and PMP inhibitor (SCOP g.4.1.1) families. Even NOE data sets for which the corresponding structure is not represented in the 95% sequence similarity filtered SCOP list yield good results, i.e. the method is able to find related structures in the database (e.g. 1d3z finds 9 relatives in the ubiquitin family in the first 10 hits). This shows the general applicability of the method. It is clear that the relative number of restraints has a profound, but not decisive effect on the hits: the number of positive hits usually increases with the number of restraints, but the relationship is more complex. According to the principles of the PRIDE-NMR method, the most important factor is how well the restraints represent the structure, which generally, but not always improves with the increasing number of NOE restraints (compare 1g6j with 1p1a). Note that other types of restraints might also be used simultaneously for NMR structure determination, thus our results do not directly reflect the 'quality' of the database structures.

Similar results were obtained for the SH3 domain and PMP inhibitor families (see the PRIDE-NMR web site) with only one protein not yielding any positive hits among the first 10 and displaying the general but not exclusive correspondence between the number of positive hits and average number of restraints per residue.

Table 1. Summary of PRIDE-NMR search with the ubiquitin test set

PDB ID	Length	Average number of restraints per residue[a]	Represented in our database/found itself	Number of relatives found[b]	
				F	S
1d3z	76	3.29	No/No	9	9
1g6j	76	2.62	No/No	2	2
1p1a	85	2.09	Yes/Yes	5	5
1m94	93	1.66	Yes/No	1	2
1mg8	78	0.53	Yes/No	1	1

Results for the first 10 hits using score weighting (third power: PRIDE-NMR_W_3) are reported (no other length filtering was used), using averaged scores for cutoff distances of 5 and 6 Å.

[a]Only unambiguously intrabackbone restraints are considered.

[b]Self hits are included; F: at the SCOP Family level (d.15.1.1); S: at the SCOP Superfamily level (d.15.1). There is only one case (1m94) where a positive hit in the ubiquitin-like superfamily but not in the more restricted ubiquitin-related family is found.

For a more comprehensive test, a set of another 40 proteins with available NMR distance restraints, covering a wide range of folds (each classified differently at the fourth level of the SCOP hierarchy, representing 40 families and 37 superfamilies, for a complete list of the domains and results, see the PRIDE-NMR web site) was selected. These domains all have relatives (domains in the same family) in the database, have an average number of intrabackbone restraints per residue above 1, and are of varying lengths (24–182 residues). A test with criteria similar to that performed by Novotny *et al*. (2003) to assess protein fold comparison servers was performed: hits in the same Superfamily (third level of classification) as the query were considered positive and the first 100 hits were monitored excluding self-hits. However, we note that in our case the usual procedures to asses server performance should be used with care

as the data in our server database and the input NMR-based distances do not correspond to each other on a one-to-one basis (i.e. the input data set is not a subset of the server database as could be for, e.g. a protein fold comparison method).

Best results were obtained using a cutoff distance of 5 Å or the averaged scores calculated for 5 and 6 Å (Table 2): in these cases, the method resulted at least one positive hit within the first 100 for 100% and 97% of the queries, respectively. This success rate is comparable to those reported for the best protein comparison servers (using a different set of proteins, Novotny *et al*., 2003). However, having a single positive hit among the first 100 is clearly not sufficient for quick structure estimation. Again, the quality of NMR distance data (which cannot be expected to be uniform in our data set) is prevalent, and thus, the exact position of the first true positive (if known) in the hit list can be used to assess the completeness of NOE data (see below).

Tests were also run using randomly truncated data sets: the number of distances was decreased by reducing the contents of randomly selected bins until the desired percentage of data remained. This procedure was applied both to the back-calculated (from 3D domain structures) and NOE data sets using the same 40 domains as above and repeating the random truncation 10 times for each data set (Fig. 2).

Table 2. Performance of the PRIDE-NMR method with the 40-protein test set

Distance cutoff(s)	Percentage of queries yielding positive hits in the:		
	first 5 hits	first 10 hits	first 100 hits
5	*57*	*62*	*100*
6	42	53	85
7	25	38	82
5,6	*60*	*62*	*97*
6,7	38	60	93
5,6,7	55	65	95

Results obtained for the 40-protein test set with score weighting (third power: PRIDE-NMR_W3) are reported (no other length filtering was used). Distance cutoffs are given in Å, where multiple distances are given, the average score obtained with the individual distances was used for ranking the hits. Cutoffs with the best results are italicized.

Interestingly, a database search with the back-calculated distance distributions yields considerably worse results than using NOE-based data even without truncation (~88% positive hits compared to 97–100%; Fig. 2). This finding is especially surprising given that the back-calculated distance distributions (calculated using a distance cutoff of 6 Å) contain about an order of magnitude more data than the NMR restraint sets. [We note that this can be regarded as a respectable performance compared to the best protein fold comparison servers (Novotny *et al*., 2003) and does not point to serious classification errors.] In our view, this may partly be explained by the insufficient ability of H–H distance distributions to represent the folds (e.g. the PRIDE method uses 28 sets of Cα–Cα distance distributions; Carugo and Pongor, 2002; Gáspári *et al*., 2005). On the other hand, it seems that NOE data sets, although sparse because of experimental errors and protein dynamics, represent the 'quintessence' a protein fold (i.e. they incorporate the most important H–H distances). The improvement caused by multiple cutoffs is in conceptual agreement with the dynamic nature of proteins and with the notion that a single conformer cannot fulfill all restraints simultaneously (Lindorff-Larsen *et al*., 2004). The scarcity of NMR distance data also underlines the importance of sophisticated structure calculation methods with reliable force fields to obtain high quality, biologically relevant structural models (Richter *et al*., 2007).

In summary, PRIDE-NMR is a simple method yielding results well within a minute. We foresee the following application areas:

- Obtaining a first guess about the fold before structure calculation, complementing sequence and chemical shift information.
- Estimating the completeness of NOE data (i.e. whether or not the available restraints are sufficient for structure determination) in cases when the target structure is related to a known one.
- Detecting of errors in resonance assignment if the target structure has known homolog(s).

With PRIDE-NMR, these checks can be performed routinely and multiple times during structure determination, allowing avoidance of futile calculations with erroneous or incomplete NOE data sets.

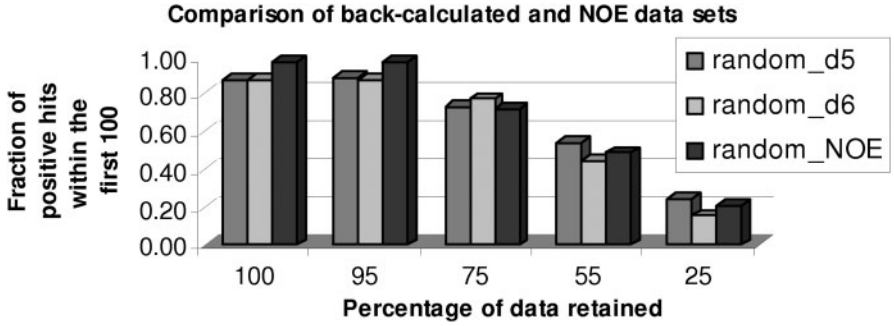

Fig. 2. Performance of the method with randomly truncated back-calculated (cutoff distances 5 and 6 Å) and NOE data sets (averaged scores for cutoffs of 5 and 6 Å). Columns represent the fraction of the query data set yielding at least one positive hit in the 100 (compare to data in Table 2).

ACKNOWLEDGEMENTS

Funding Grants from the Hungarian Scientific Research Fund (OTKA F68079, TS049812, T046994) and the International Centre for Genetic Engineering and Biotechnology (Hun-04-03), as well as a János Bolyai Research Fellowship (to Z.G.) are acknowledged.

Conflict of Interest: none declared.

REFERENCES

Annila,A. *et al*. (1999) Recognition of protein folds via dipolar couplings. *J. Biomol. NMR*, **14**, 223–230.

Brünger,A.T. *et al*. (1998) Crystallography & NMR System: a new software suite for macromolecular structure determination. *Acta Crystallogr. D Biol. Crystallogr*., **54**, 905–921.

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.

Cavalli,A. *et al*. (2007) Protein structure determination from NMR chemical shifts. *Proc. Natl Acad. Sci. USA*, **104**, 9615–9620.

Delaglio,F. *et al*. (2000) Protein structure determination using molecular fragment replacement and NMR dipolar couplings. *J. Am. Chem. Soc*., **122**, 2142–2143.

Gáspári,Z. *et al*. (2005) Efficient recognition of folds in protein 3D structures by the improved PRIDE algorithm. *Bioinformatics*, **21**, 3322–3323.

Herrmann,T. *et al*. (2002) Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS. *J. Biomol. NMR*, **24**, 171–189.

Kaminski,G.A. *et al*. (2001) Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem*., **105**, 6474–6487.

Lindorff-Larsen,K. *et al*. (2004) Simultaneous determination of protein structure and dynamics. *Nature*, **433**, 128–132.

Murzin,A.G. *et al*. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol*., **247**, 536–540.

Novotny,M. *et al*. (2003) Evaluation of protein fold comparison servers. *Proteins*, **233**, 260–270.

Richter,B. *et al*. (2007) The MUMO (minimal under-restraining minimal over-restraining) method for the determination of native state ensembles of proteins. *J. Biomol. NMR*, **37**, 117–135.

Rieping,W. *et al*. (2007) ARIA2: automated NOE assignment and data integration in NMR structure calculation. *Bioinformatics*, **23**, 381–382.

Sonego,P. *et al*. (2007) A protein classification benchmark collection for machine learning. *Nucleic Acids Res*., **35**, D232–D236.

van der Spoel,D. *et al*. (2005) GROMACS: fast, flexible and free. *J. Comput. Chem*., **26**, 1701–1718.