Probing Dynamic Protein Ensembles with Atomic Proximity Measures

Zoltán Gáspári[1,]*, Annamária F. Ángyán[1,2], Somdutta Dhir[3], Dino Franklin[4], András Perczel[1], Alessandro Pintar[3] and Sándor Pongor[3,]*

[1]*Laboratory of Structural Chemistry and Biology, Institute of Chemistry, Eötvös Loránd University, Budapest, Hungary;* [2]*ELTE-HAS Protein Modeling Group, Budapest, Hungary,* [3]*Protein Structure and Bioinformatics, ICGEB, Trieste, Italy,* [4]*Computer Science Department, Campus de Catalão, Federal University of Goiás, Catalão, Brazil*

Abstract: The emerging role of internal dynamics in protein fold and function requires new avenues of structure analysis. We analyzed the dynamically restrained conformational ensemble of ubiquitin generated from residual dipolar coupling data, in terms of protruding and buried atoms as well as interatomic distances, using four proximity-based algorithms, CX, DPX, PRIDE and PRIDE-NMR (http://hydra.icgeb.trieste.it/protein/). We found that Ubiquitin, this relatively rigid molecule has a highly diverse dynamic ensemble. The environment of protruding atoms is highly variable across conformers, on the other hand, only a part of buried atoms tends to fluctuate. The variability of the ensemble cautions against the use of single conformers when explaining functional phenomena. We also give a detailed evaluation of PRIDE-NMR on a wide dataset and discuss its usage in the light of the features of available NMR distance restraint sets in public databases.

Keywords: Protein structural ensemble, protein NMR, atom depth, solvent exposition, protein structure comparison.

INTRODUCTION

Understanding the principles of protein structures is essential to gain insight into biomolecular processes. Enzymatic conversions, protein-protein interactions are all based on intermolecular recognition dictated by the shapes and interacting surface groups of the partners. Recently, the role of internal dynamics manifested as structural fluctuations is gradually coming into focus [1-3]. The inherently dynamic nature of proteins requires the introduction of new ways of structure representation, fold identification and comparison.

Protein folds are manifested as a set of atomic coordinates in the three-dimensional space. However, all-atom representations of proteins are not really efficient for most practical purposes such as protein visualization or structure comparison. Moreover, the inherently dynamic nature of proteins means that the set of atomic coordinates is continuously changing. Thus, simplified representations capturing the essence of protein folds or specific aspects of structures have been developed. The most common representations are based on the topology of secondary structure elements and atom-atom distances. There are measures that assign a single number to protein folds based on their residue-residue packing properties. These measures, contact order [4] and long-range order [5], have been shown to correlate with protein folding rates.

The methods reviewed here, available as free web services at ICGEB (http://hydra.icgeb.trieste.it, [6]), make use of close or closest distances to capture features of folds or finding structural relatives. CX and DPX assign a parameter to each of the protein atoms based on their proximity to solvent, whereas PRIDE and PRIDE-NMR use distributions of (typically short) interatomic distances to represent folds enabling rapid scan of structural databases. For the most recently described method, PRIDE-NMR, we give a detailed evaluation of its performance on a wide dataset and discuss its performance and usage in the light of the quality and completeness of available NMR distance restraint sets.

Considering the dynamic nature of proteins, we describe the applicability of all four methods to dynamic conformational ensembles. Structural heterogeneity in such ensembles stems from the solution-state internal dynamics of the protein in a range of time scales [7-9]. As a test case, we use the recently published dynamic ubiquitin ensemble calculated from RDC data and covering dynamics up to the millisecond range (PDB ID 2K39, containing 116 conformers) [3]. This ensemble has been shown to include conformers close to those observed in known structures of ubiquitin complexes with other proteins, thus sufficiently sampling the conformational space accessible to ubiquitin.

DETAILED EVALUATION OF PRIDE-NMR

PRIDE-NMR [10] is a conceptually simple and fast method that is able to recognize protein folds compatible with a given set of NOE distance data. PRIDE-NMR is not a protein fold comparison method, although shares clear conceptual similarities to those, especially to PRIDE (Probability of IDEntity, [11, 12], a fast approach based on the comparison of Hα-Hα distance distributions in protein structures. PRIDE-NMR is based on the comparison of distributions of short interproton distances observable as NOE peaks in NMR spectra and easily obtainable from known 3D protein structures. The number of NOE-derived distance restraints or close H-H pairs is represented in a histogram with bins cor-

*Address correspondence to these authors at the Laboratory of Structural Chemistry and Biology, Institute of Chemistry, Eötvös Loránd University, Budapest, Hungary; Tel: +36-1-3722500; Fax: +36-1-3722620
E-mail: szpari@chem.elte.hu; and Protein Structure and Bioinformatics, ICGEB, Trieste, Italy; Tel: +39-040-3757300; Fax: +39-040-3757341
E-mail: pongor@icgeb.org

1389-2037/10 $55.00+.00

responding to the sequential separation of the participating residues. To render the algorithm largely independent of the sequences of the proteins, only HN, Hα and Hβ protons are used for compiling the distributions.

The histograms are then compared using contingency analysis Fig. (**1**) similar to the PRIDE method [11]. To enhance the specificity of results, a simple weighting of the scores was also introduced based on the lengths of the query and hit proteins Eq. (1):

$$PRIDE-NMR\_W_x = PRIDE-NMR * \left(\frac{length\_of\_shorter\_protein}{length\_of\_longer\_protein}\right)^x \quad \text{Eq. (1)}$$

Where *PRIDE-NMR* is the raw score calculated with the contingency test, *PRIDE-NMR_W$_x$* is the weighted score according to exponent *x* where *x* can be 1, 2 or 3 in the current implementation. The weighted score is always less or equal to the original one, ensured by choosing the length of the longer protein as denominator, regardless of whether it corresponds to the query dataset or to one in the database.

The PRIDE-NMR method was initially evaluated on a set of 40 proteins chosen as a representative subset of the SCOP database and having a restraint/residue ratio greater than 1.0 [10]. In order to yield a more realistic picture of the performance of the approach with typical NMR datasets, we performed a more comprehensive evaluation using as many protein structures as possible at the time of analysis. The protein test set was compiled as all available single-domain NMR structures in the PDB [13] as of 24 September 2007 classified in the SCOP database [14], having at least one homologue in SCOP at the family level and with deposited NMR distance restraint set in X-PLOR format (Fig. **2**).

We have performed the search and evaluation as described previously [10]. The database is based on the 95% homology-filtered subset of SCOP and structures were taken from the ASTRAL database [15]. We carefully analyzed the reasons of failure where the method could not identify structural relatives of the protein corresponding to the query dataset, defined as finding hits in the same SCOP family as the query.

Performance of the PRIDE-NMR Approach on a Large Set of Structures

In accordance with our previous test, the test was positive if when the first N hits contained a structural relative of the query at the SCOP superfamily level [10], where the cases of N=5 and N=100 were investigated. The performance of the PRIDE-NMR approach on a large data set is somewhat lower than reported for a limited set of structures. However, the performance of 75% is still comparable to those of the better performing protein fold comparison servers [16]. Direct comparison is impossible because of the differences in the test data sets, as in PRIDE-NMR the query dataset is obtained differently than those in the database, i.e. experimental vs. structure-derived distances are used, whereas protein fold comparison servers use structure-derived data both for storage and query. Thus, for PRIDE-NMR cases where the restraint set is capable to retrieve its corresponding PDB entry could also be regarded positives. However, to conform to conventions in evaluating fold comparison methods, self hits, i.e. where the query and the hit belongs to the same PDB ID were excluded from the analysis.

As reported earlier [10], best results are obtained using a distance cutoff of 5 Å for back-calculating short interproton

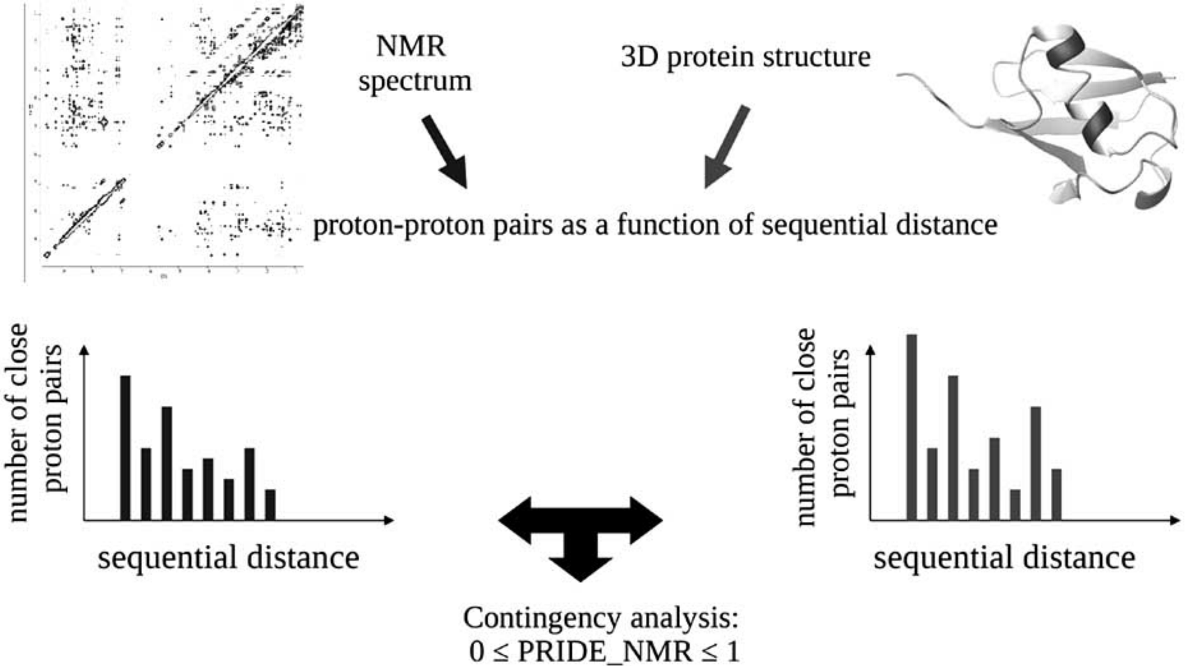

NMR spectrum
3D protein structure
proton-proton pairs as a function of sequential distance
number of close proton pairs
sequential distance
number of close proton pairs
sequential distance
Contingency analysis:
0 ≤ PRIDE_NMR ≤ 1

Fig. (1). Schematic outline of the PRIDE-NMR method. PRIDE-NMR uses statistical comparison of distance distributions from NOE data measured experimentally and back-calculated from structures. The distance distribution represents the number of close proton pairs vs. the sequential distance. The resulting probability is the PRIDE score which can be weighted according to Eq. (**1**). (The number of pairs and the sequential distance are dimensionless quantities.)

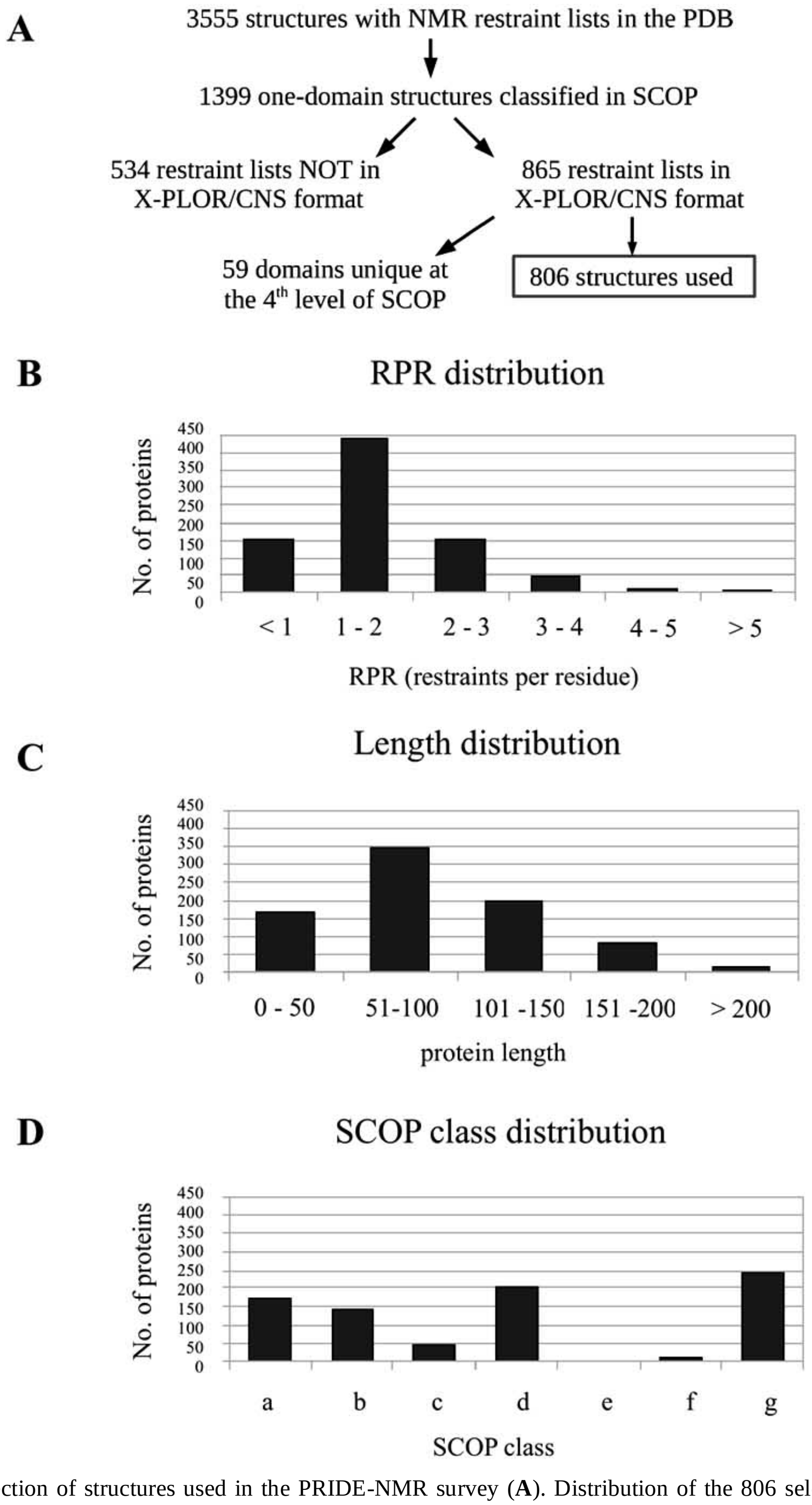


Fig. (2). Scheme of the selection of structures used in the PRIDE-NMR survey (**A**). Distribution of the 806 selected proteins according to (intrabackbone) restraint per residue ratios (**B**), length (**C**), and SCOP class (**D**). The SCOP class categories are the following: **a**) all-α **b**) all-β **c**) α/β **d**) α+β **e**) multidomain proteins **f**) membrane and cell surface proteins **g**) small proteins

distances from 3D structures, or when the PRIDE-NMR scores obtained for the 5 and 6 Å or the 5,6 and 7 Å cutoffs are averaged (Table **1**). This indicates that, not surprisingly, NOEs corresponding to 5 Å or shorter distances dominate the deposited restraint lists.

Investigating the cases not yielding positive hits in our PRIDE-NMR survey, we could identify three common reasons why NOE datasets are unable to represent the fold properly in our approach:

Table 1. PRIDE-NMR Results for the 806-Membered Test Set. Self Hits (i.e. when the Hit has The Same PDB ID as the Query) were Excluded From The Analysis. The Distances in Ångstrøms Refer to the Distance Cutoff(s) Used (Multiple Numbers Mean that Both Cutoffs were Used and The Scores Averaged). Weighting Refers to The Scores Calculated as Shown in Eq. (1)

Distance Cutoff		**d=5 Å**		**d=6 Å**		**d=7 Å**		**d=5, 6 Å**		**d=6, 7 Å**		**d=5, 6, 7 Å**	
Weighting		**First 5**	**First 100**	**First 5**	**First 100**	**First 5**	**First 100**	**First 5**	**First 100**	**First 5**	**First 100**	**First 5**	**First 100**
W0	positive hit %	**17.62**	**49.38**	**10.79**	**36.60**	**8.19**	**30.40**	**17.62**	**46.40**	**10.05**	**36.48**	**13.52**	**42.80**
	avg. no. of positive hits	0.33	2.46	0.20	1.86	0.17	1.60	0.32	2.39	0.20	1.87	0.26	2.22
W1	positive hit %	**36.85**	**67.74**	**20.60**	**50.99**	**15.63**	**43.92**	**36.72**	**67.74**	**21.34**	**55.46**	**34.12**	**66.38**
	avg. no. of positive hits	0.75	4.36	0.41	2.84	0.30	2.37	0.77	4.33	0.43	3.26	0.69	4.11
W2	positive hit %	**39.08**	**70.72**	**23.45**	**57.32**	**18.24**	**51.36**	**39.83**	**71.96**	**24.81**	**63.15**	**36.97**	**71.59**
	avg. no. of positive hits	0.81	5.07	0.46	3.33	0.35	2.91	0.83	5.13	0.51	3.94	0.77	5.01
W3	positive hit %	**43.67**	**75.06**	**27.79**	**63.65**	**19.73**	**57.57**	**42.06**	**75.31**	**28.78**	**66.38**	**40.45**	**75.31**
	avg. no. of positive hits	0.88	5.67	0.53	3.92	0.39	3.57	0.88	5.75	0.58	4.63	0.83	5.68

- Restraint lists may contain only limited data, e.g. number of intrabackbone NOEs < 10 or representing less than 5 sequential distances (bins). It should be noted that structure determination may use a wealth of other NMR information besides NOE data, thus scarceness of NOE data does not necessarily reflect a structure of poor accuracy and may be biologically relevant anyway.
- There are some highly biased distance distributions: if the majority of the data fall into one or two bins, the combination of bins to ensure that none of them contains less than 5% of the data results in erroneously low probability values upon comparisons to other distributions. This is observed for structures with helical segments and few long-range NOEs. By default, PRIDE-NMR uses the minimum sequential distance of 3 residues (NOEs between residues *i*-*i*+3 are used), this can be adjusted to 4 or 5 to avoid computational problems with helical structures.
- Several structures contained suspiciously high number of intrabackbone restraints per residue and these were also poorly performing in the tests.
- There can be some other reasons, e.g. a structure atypical for its SCOP family in some respect (e.g. chain length) might also fail in our approach with weighting, and in this case the unweighted results might be relevant.

Examples of Poorly Performing Structures

We have selected three structures for which PRIDE-NMR fails to identify structural relatives for various reasons Fig. (**3**).

- **1P88:** All related domains in SCOP are much longer, over 400 residues long, thus, PRIDE-NMR's weighting scheme, ensuring that only proteins of similar size yield top results, diminishes its performance. Unweighted results are realistic and representative (first 3 hits are in the same SCOP class), proving the usability of the PRIDE-NMR concept itself and emphasizing the importance of separate parameterization in special cases.
- **1VND** [17]**:** Strangely, all possible H-H distances are represented in the restraint file, those not corresponding to observed NOEs with an artificially high distance limit (99 Å). As PRIDE-NMR does not check for the distance represented by the restraints, assuming the only distances corresponding to observed NOEs are reported, this results in an unrealistically high RPR (restraint per residue) ratio and completely irrelevant results.
- **1SP2** [18]: In this short protein there are relatively few NOEs, representing too few NOE categories for a meaningful comparison, as contingency analysis fails due to the low number of bins to compare. This is a common problem for proteins with scarce NOE distance data, thus care should be taken when using structures with RPR < 1.

Practical Aspects of PRIDE-NMR

The PRIDE-NMR approach is capable of relating NOE distance data to known protein folds. Although it is not a protein fold comparison server, methods for testing its performance are adapted from those applied to fold comparison servers, as they show the closest similarity to its concept and are familiar to the structural biology community.

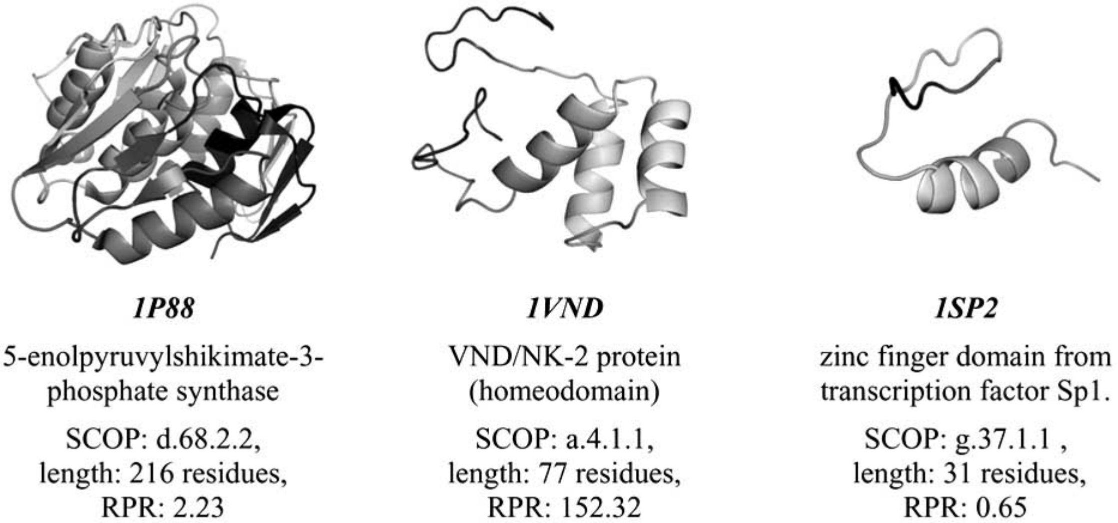


Fig. (3). Selected examples for which the PRIDE-NMR search fails. See text for detailed explanation of the reason of failure (RPR: intra-backbone restraints per residue)

In the present review, using the widest possible dataset we were able to show the performance of the method on data commonly available and thus we are able to show the weaknesses of both deposited NMR NOE data and the capability of the PRIDE-NMR approach, and can highlight the practical aspects of using the method, which can be summarized as:

- Scarceness of NOE data seriously hampers successful application of the method. The (intrabackbone) NOE / residue ratio should best be above 1.
- Biased restraint distribution can also be a serious drawback. If there are too few NOE categories, the method fails. Special care should be taken for helical proteins, where the high number of *i-i+3* NOEs can yield unspecific results. For these proteins, setting the minimal sequential distance to 4 is recommended.
- For proteins of atypical size among their relatives, ignoring length weighting may improve the significance of the results.
- For a dissimilar set of protein structures, the selected one in our database might not be the one yielding the highest score with the input dataset. This is a consequence of structural heterogeneity in deposited ensembles, an issue discussed in more detail below.

In general, running the search with a few parameter combinations is advised, as the speed of the method allows this to be done in a few minutes and can improve the results.

Protein Fold Recognition and Information Content of NOE Distances

^{1}H-^{1}H NOEs carry information about the atomic-level structure of a protein sufficient for high-quality structure calculations. Therefore, an initial assumption during the development of the PRIDE-NMR method was that proton-proton distances available from experiments might be sufficient for unambiguous 3D fold identification from a database similarly to the Cα-Cα distance distributions as used in the PRIDE approach [11, 19]. The original PRIDE approach uses separate histograms for 28 sequential distances (from 3 to 30), each with bins of 1 Å width [11]. Indeed, replacing the Cα coordinates of a protein with Hα ones by simple tinkering of PDB files yields practically the same results as using the original Cα coordinates (data not shown).

However, Hα-Hα pairs in experimentally derived distance restraint lists are so scarce that they carry practically no information about the fold when used alone. A probabilistic approach using information from other restraints from the same residue pair to estimate the Cα-Cα distance also failed because of the high uncertainty in the resulting distance lists (not shown). Finally, we chose to represent the data in a single histogram as a function of sequential distance and without separating NOEs between different atom pairs, i.e. all NOEs between any pairs of NH, Hα and Hβ atoms are included. This approach yields enough data to uniquely represent different folds of similar chain lengths. To reduce the potential of relating the distance restraint set of e.g. a 50-residue protein to a 150-residue structure, length weighting was introduced (see Eq. (1) above). We note that the histograms contain still only about 10% of the data obtained by back-calculating the ^{1}H-^{1}H distances from the 3D coordinates [10]. This means that nearly 90% of the close proton-proton contacts, at least when considering NH, Hα and Hβ atoms, remains 'invisible' to NMR measurements for various reasons inevitably including the internal dynamics of proteins. Nevertheless, the NMR-derived data performed slightly better in the fold recognition tests [10] suggesting that NMR is able to capture the 'essence' of protein folds in terms of ^{1}H-^{1}H distances.

PROXIMITY-BASED MEASURES AND DYNAMIC PROTEIN STRUCTURAL ENSEMBLES

CX [20] is a fast algorithm to identify solvent-exposed protruding groups in protein structures. Briefly, it analyzes the number of protein atoms in a sphere around the selected one and yields the ratio of the volume occupied by atoms outside and inside the protein. The CX values are output as B-factors in a PDB format file, rendering visualization straightforward. Identifying protruding atoms is expected to yield valuable information on residues potentially involved in interactions and modifications as shown for trypsin cleavage sites of different proteins [20]. Calculating the CX values for all non-hydrogen atoms of the 2K39 ubiquitin ensemble revealed that there is a strong correlation (R=0.93 calculated by omitting all CX values with standard deviation of zero) between the average CX value over the ensemble at its standard deviation Fig. (**4**). Interestingly, the average correlation coefficient between CX values calculated for single structures and either their average or standard deviation over the ensemble is weaker (R=0.82 and 0.75, respectively). This means that higher CX value means greater changes in atomic environment during conformational dynamics, an observation in line with previous notions on accessibility measures and crystallographic B-factors [20]. In the dynamic ubiquitin ensemble CX values successfully identify the two most common lysines (Lys48 and Lys63) linked to Gly76 in polyubiquitin chains [21]. Lys11 (as practically all lysines) is also involved in polyubiquitinylation [21] and Arg54 is involved in initial recognition by the E1 enzyme [22]. Thr9 is a potential phosphorylation site by PKC and CKII according to the KinasePhos server [23]. Residues with high CX values also include those in the flexible C-terminal tail, most of which are also involved in molecular recognition processes [21]. Naturally, as CX is a tool to identify protruding atoms, it cannot be expected by definition to be capable to identify all interaction sites such as grooves on the protein surface.

The DPX algorithm measures the depth of atoms in protein structures [24, 25] as their distance to the closest (proximal) solvent-exposed atom. DPX thus probes the protein core and is indicative of fold-stabilizing residues making key intramolecular contacts. Atom depth correlates with residue hydrophobicity and secondary structure [24]. Groups closer to the surface, i.e. with relatively low DPX value could be identified as targets for posttranslational modifications [24]. We also note that buried residues show different conformational preferences than solvent-exposed ones and their conformation can much less reliably be predicted using data from free peptide models [26]. Our results on the 116-membered dynamic ubiquitin ensemble reveal that there is no significant correlation between the average DPX values of heavy atoms and their standard deviation (R=0.59 for atoms with nonzero DPX value). While for atoms closer to the surface there seems to be some tendency as higher average DPX means higher deviations, this trend ameliorates at around DPX values of 0.5 Å, above which high average can be associated with zero or high deviation Fig. (**5**). This means that structural variations affect the protein core in a much more complex manner than surface residues. The first 10 residues bearing the atoms with largest DPX values are shown in Fig. (**5**). This visualization does not account for the number of atoms with high DPX values per residue. Nevertheless, it is interesting to note that a designed hydrophobic core variant of ubiquitin which turned out to have essentially the same structure as the wild-type one [27] bears 7 mutations from which only 3 (Ile4, Val5 and Val26) are among the residues with deepest atoms identified in our analysis. Thus, identifying and targeting structure-stabilizing residues is a fairly complex issue as judged by the diversity of published relevant methods [28-30]. We believe that consideration of the internal dynamics of proteins and its capability of changing atomic depth and the network of intramolecular interactions should be taken into account when identifying key atoms in structure stabilization.

PRIDE is a protein fold comparison method using Cα-Cα distance distributions for each sequential distance from 3 to 30 residues. Its original version [11] uses 28 histograms, one for each sequential distance, to represent distance distributions. These histograms are compared using contingency analysis in a pairwise manner for the two structures investigated. A later version of the approach (PRIDE2, [19]) uses continuous distributions and Kolmogorov-Smirnov test for

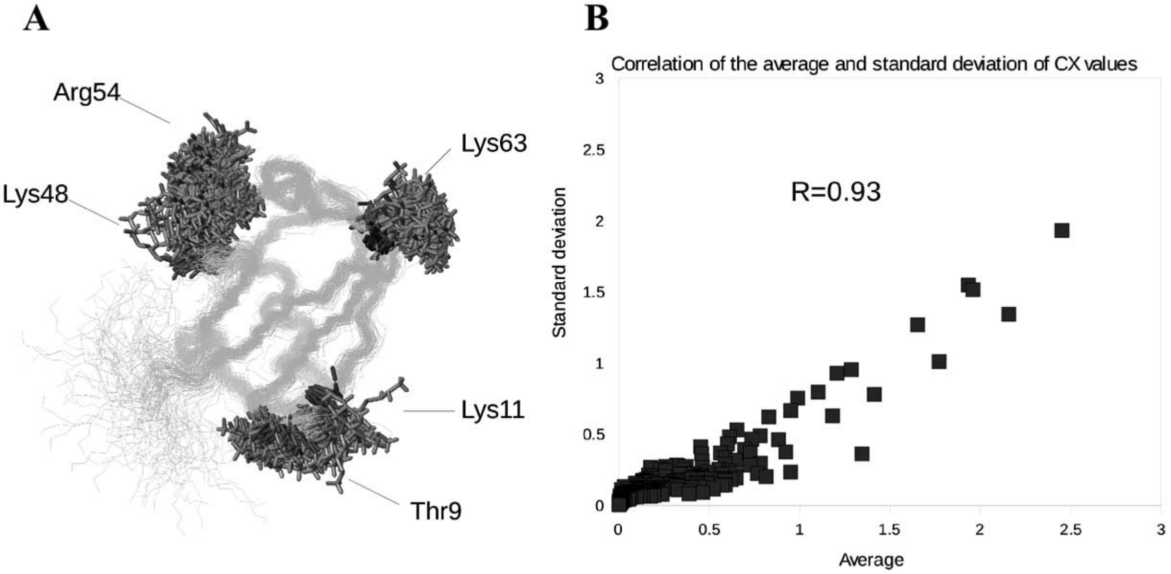


Fig. (4). Residues with high CX values in the RDC-derived dynamic ubiquitin ensemble (those in the flexible C-terminus are not highlighted) (**A**). Correlation of the average and the standard deviation of CX values of atoms in the dynamic ensemble (atoms with exactly zero standard deviation are omitted) (**B**).

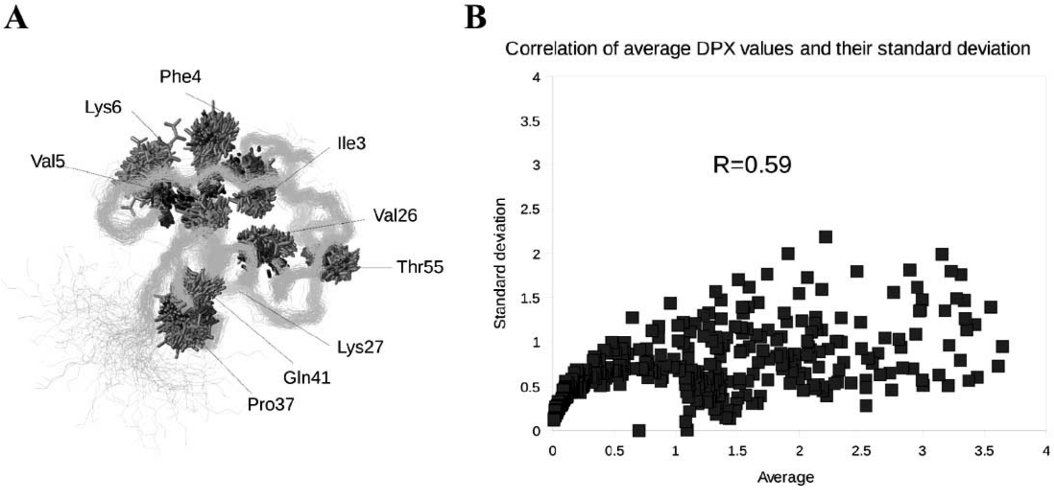
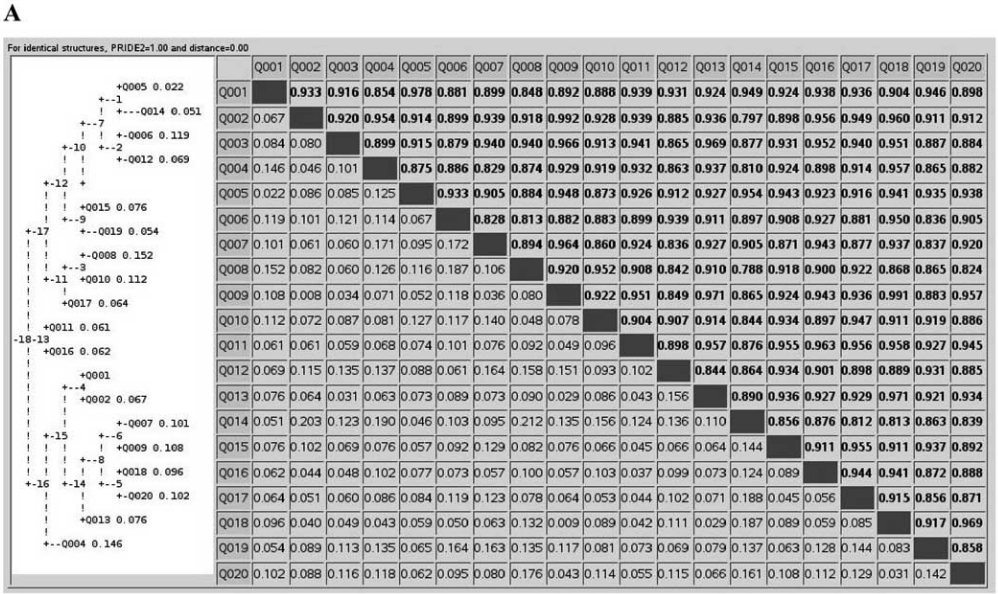


Fig. (5). Residues with highest atomic DPX values in the RDC-derived dynamic ubiquitin ensemble (**A**). Correlation of the average and the standard deviation of DPX values of atoms in the dynamic ensemble (atoms with zero average DPX are omitted) (**B**).



For identical structures, PRIDE2=1.00 and distance=0.00

```
                 +Q005 0.022
              +--1
              !  +---Q014 0.051
           +--7
           !  !  +-Q006 0.119
        +-10  +--2
        !  !     +-Q012 0.069
        !  !
     +-12  +
     !  !
     !  !  +Q015 0.076
     !  +--9
  +-17     +--Q019 0.054
  !  !
  !  !     +-Q008 0.152
  !  !  +--3
  !  +-11  +Q010 0.112
  !     !
  !     +Q017 0.064
  !
  !  +Q011 0.061
-18-13
  !  +Q016 0.062
  !
  !        +Q001
  !     +--4
  !     !  +Q002 0.067
  !     !
  !     !        +-Q007 0.101
  !  +-15     +--6
  !  !  !     !  +Q009 0.108
  !  !  !  +--8
  !  !  !  !  !  +Q018 0.096
  +-16  +-14  +--5
     !     !     +-Q020 0.102
     !     !
     !     +Q013 0.076
     !
     +--Q004 0.146
```

	Q001	Q002	Q003	Q004	Q005	Q006	Q007	Q008	Q009	Q010	Q011	Q012	Q013	Q014	Q015	Q016	Q017	Q018	Q019	Q020
Q001	■	0.933	0.916	0.854	0.978	0.881	0.899	0.848	0.892	0.888	0.939	0.931	0.924	0.949	0.924	0.938	0.936	0.904	0.946	0.898
Q002	0.067	■	0.920	0.954	0.914	0.899	0.939	0.918	0.992	0.928	0.939	0.885	0.936	0.797	0.898	0.956	0.949	0.960	0.911	0.912
Q003	0.084	0.080	■	0.899	0.915	0.879	0.940	0.940	0.966	0.913	0.941	0.865	0.969	0.877	0.931	0.952	0.940	0.951	0.887	0.884
Q004	0.146	0.046	0.101	■	0.875	0.886	0.829	0.874	0.929	0.919	0.932	0.863	0.937	0.810	0.924	0.898	0.914	0.957	0.865	0.882
Q005	0.022	0.086	0.085	0.125	■	0.933	0.905	0.884	0.948	0.873	0.926	0.912	0.927	0.954	0.943	0.923	0.916	0.941	0.935	0.938
Q006	0.119	0.101	0.121	0.114	0.067	■	0.828	0.813	0.882	0.883	0.899	0.939	0.911	0.897	0.908	0.927	0.881	0.950	0.836	0.905
Q007	0.101	0.061	0.060	0.171	0.095	0.172	■	0.894	0.964	0.860	0.924	0.836	0.927	0.905	0.871	0.943	0.877	0.937	0.837	0.920
Q008	0.152	0.082	0.060	0.126	0.116	0.187	0.106	■	0.920	0.952	0.908	0.842	0.910	0.788	0.918	0.900	0.922	0.868	0.865	0.824
Q009	0.108	0.008	0.034	0.071	0.052	0.118	0.036	0.080	■	0.922	0.951	0.849	0.971	0.865	0.924	0.943	0.936	0.991	0.883	0.957
Q010	0.112	0.072	0.087	0.081	0.127	0.117	0.140	0.048	0.078	■	0.904	0.907	0.914	0.844	0.934	0.897	0.947	0.911	0.919	0.886
Q011	0.061	0.061	0.059	0.068	0.074	0.101	0.076	0.092	0.049	0.096	■	0.898	0.957	0.876	0.955	0.963	0.956	0.958	0.927	0.945
Q012	0.069	0.115	0.135	0.137	0.088	0.061	0.164	0.158	0.151	0.093	0.102	■	0.844	0.864	0.934	0.901	0.898	0.889	0.931	0.885
Q013	0.076	0.064	0.031	0.063	0.073	0.089	0.073	0.090	0.029	0.086	0.043	0.156	■	0.890	0.936	0.927	0.929	0.971	0.921	0.934
Q014	0.051	0.203	0.123	0.190	0.046	0.103	0.095	0.212	0.135	0.156	0.124	0.136	0.110	■	0.856	0.876	0.812	0.813	0.863	0.839
Q015	0.076	0.102	0.069	0.076	0.057	0.092	0.129	0.082	0.076	0.066	0.045	0.066	0.064	0.144	■	0.911	0.955	0.911	0.937	0.892
Q016	0.062	0.044	0.048	0.102	0.077	0.073	0.057	0.100	0.057	0.103	0.037	0.099	0.073	0.124	0.089	■	0.944	0.941	0.872	0.888
Q017	0.064	0.051	0.060	0.086	0.084	0.119	0.123	0.078	0.064	0.053	0.044	0.102	0.071	0.188	0.045	0.056	■	0.915	0.856	0.871
Q018	0.096	0.040	0.049	0.043	0.059	0.050	0.063	0.132	0.009	0.089	0.042	0.111	0.029	0.187	0.089	0.059	0.085	■	0.917	0.969
Q019	0.054	0.089	0.113	0.135	0.065	0.164	0.163	0.135	0.117	0.081	0.073	0.069	0.079	0.137	0.063	0.128	0.144	0.083	■	0.858
Q020	0.102	0.088	0.116	0.118	0.062	0.095	0.080	0.176	0.043	0.114	0.055	0.115	0.066	0.161	0.108	0.112	0.129	0.031	0.142	■

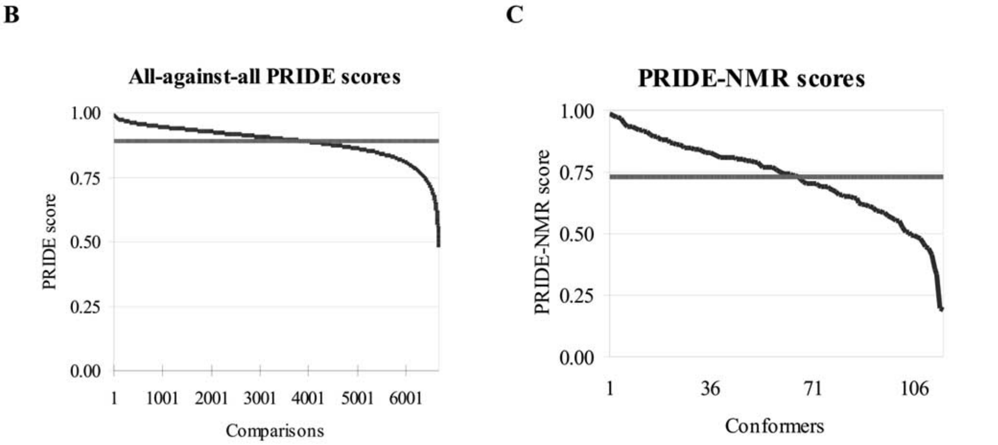


Fig (6). PRIDE and PRIDE-NMR as measures of ensemble heterogeneity. Output of the PRIDE2 server for clustering the first 20 conformers for the dynamic ubiquitin ensemble (2K39) (**A**). A neighbor-joining tree along with a matrix containing the PRIDE scores (upper triangle) and the 1-PRIDE distances (lower triangle) are shown. All-against-all PRIDE2 scores for all 116 structures in the same ensemble (6670 comparisons) (**B**) and PRIDE-NMR scores for all 116 members as calculated using the available NOE restraint set for the structure 1D3Z [31] (**C**) In (**B**) and (**C**), the average score is represented by a horizontal line.

their comparison. PRIDE-NMR is a conceptually related tool capable of relating protein structures to NMR-derived distance restraint sets (see above). Such sets contain hydrogen-hydrogen distances in the maximal range of 5-6 Å. Both PRIDE and PRIDE-NMR are quite fast and can be used as a nearest-neighbor classifier when interpreting results of database searches.

Here we use these two methods to assess the heterogeneity of dynamic protein ensembles. For PRIDE, the Cluster option available in the web server can be used to relate ensemble members to each other. In the case of PRIDE-NMR an NMR NOE distance restraint list is searched against a 'database' consisting of the ensemble members.

In accordance with principal component analysis [3]., neighbor-joining clustering of PRIDE2 all-against-all scores reveals distinct conformer families within the 116-membered ensemble. Investigation of the all-against-all PRIDE scores shows considerable heterogeneity within the conformers as scores as low as 0.5 can be observed, which are generally not considered indicative of extensive similarity in a database search. Conceptually similar results were obtained with PRIDE-NMR where the score depends heavily on the conformer selected (Fig. **6**). Although the average scores are acceptably high in both cases, these results underline the importance of conformer selection for comparative structure analyses such as database searches, where protein structures are usually represented by a single conformer.

ACKNOWLEDGEMENTS

The authors are indebted to all previous coworkers who helped the earlier phases of the project and especially to O. Carugo for his fundamental role in developing the PRIDE, CX and DPX algorithms as well as to K. Vlahovicek for the WWW server applications. The work at ELTE was supported by grants from ICGEB (CRP/HUN08-03), the Hungarian Scientific Research Fund (OTKA F68079 and K72973), as well as a János Bolyai Research Fellowship and a FEBS Short-term Fellowship to Z.G.

REFERENCES

[1] Eisenmesser, E.Z.; Millet, O.; Labeikovsky, W.; Korzhnev, D.M.; Wolf-Watz, M.; Bosco, D.A.; Skalicky, J.J.; Kay, L.E.; Kern, D. Intrinsic dynamics of an enzyme underlies catalysis. *Nature,* **2005**, *438*(7064), 117-21.

[2] Henzler-Wildman, K.; Kern, D. Dynamic personalities of proteins. *Nature,* **2007**, *450*(7172), 964-72.

[3] Lange, O.F.; Lakomek, N.A.; Fares, C.; Schroder, G.F.; Walter, K.F.; Becker, S.; Meiler, J.; Grubmuller, H.; Griesinger, C.; de Groot, B.L. Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science,* **2008**, *320*(5882), 1471-5.

[4] Gillespie, B.; Plaxco, K.W. Using protein folding rates to test protein folding theories. *Annu. Rev. Biochem.,* **2004**, *73,* 837-59.

[5] Gromiha, M.M.; Selvaraj, S. Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction. *J. Mol. Biol.,* **2001**, *310*(1), 27-32.

[6] Vlahovicek, K.; Pintar, A.; Parthasarathi, L.; Carugo, O.; Pongor, S. CX, DPX and PRIDE: WWW servers for the analysis and comparison of protein 3D structures. *Nucleic Acids Res.,* **2005**, *33,* W252-4.

[7] Lindorff-Larsen, K.; Best, R.B.; Depristo, M.A.; Dobson, C.M.; Vendruscolo, M. Simultaneous determination of protein structure and dynamics. *Nature,* **2005**, *433*(7022), 128-32.

[8] Richter, B.; Gsponer, J.; Varnai, P.; Salvatella, X.; Vendruscolo, M. The MUMO (minimal under-restraining minimal over-restraining) method for the determination of native state ensembles of proteins. *J. Biomol. NMR,* **2007**, *37*(2), 117-35.

[9] Gaspari, Z.; Varnai, P.; Szappanos, B.; Perczel, A. Reconciling the lock-and-key and dynamic views of canonical serine protease inhibitor action. *FEBS Lett.,* **2010**, *584*(1), 203-6.

[10] Angyan, A.F.; Perczel, A.; Pongor, S.; Gaspari, Z. Fast protein fold estimation from NMR-derived distance restraints. *Bioinformatics,* **2008**, *24*(2), 272-5.

[11] Carugo, O.; Pongor, S. Protein fold similarity estimated by a probabilistic approach based on C(alpha)-C(alpha) distance comparison. *J. Mol. Biol.,* **2002**, *315*(4), 887-98.

[12] Vlahovicek, K.; Carugo, O.; Pongor, S. The PRIDE Server for protein three-dimensional similarity. *J. Appl. Cryst.,* **2002**, *35*(5), 648-649.

[13] Berman, H.M. The Protein Data Bank: a historical perspective. *Acta. Crystallogr. A.,* **2008**, *64*(Pt 1), 88-95.

[14] Murzin, A.G.; Brenner, S.E.; Hubbard, T.; Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.,* **1995**, *247*(4), 536-540.

[15] Chandonia, J.M.; Hon, G.; Walker, N.S.; Lo Conte, L.; Koehl, P.; Levitt, M.; Brenner, S.E. The ASTRAL Compendium in 2004. *Nucleic Acids Res.,* **2004**, *32*(*Database issue*), D189-92.

[16] Novotny, M.; Madsen, D.; Kleywegt, G.J. Evaluation of protein fold comparison servers. *Proteins,* **2004**, *54*(2), 260-270.

[17] Tsao, D.H.; Gruschus, J.M.; Wang, L.H.; Nirenberg, M.; Ferretti, J.A. The three-dimensional solution structure of the NK-2 homeodomain from Drosophila. *J. Mol. Biol.,* **1995**, *251*(2), 297-307.

[18] Narayan, V.A.; Kriwacki, R.W.; Caradonna, J.P. Structures of zinc finger domains from transcription factor Sp1. Insights into sequence-specific protein-DNA recognition. *J. Biol. Chem.,* **1997**, *272*(12), 7801-7809.

[19] Gaspari, Z.; Vlahovicek, K.; Pongor, S. Efficient recognition of folds in protein 3D structures by the improved PRIDE algorithm. *Bioinformatics,* **2005**, *21*(15), 3322-3323.

[20] Pintar, A.; Carugo, O.; Pongor, S. CX, an algorithm that identifies protruding atoms in proteins. *Bioinformatics,* **2002**, *18*(7), 980-984.

[21] Hicke, L.; Schubert, H.L.; Hill, C.P. Ubiquitin-binding domains. *Nat. Rev. Mol. Cell Biol.,* **2005**, *6*(8), 610-621.

[22] Burch, T.J.; Haas, A.L. Site-directed mutagenesis of ubiquitin. Differential roles for arginine in the interaction with ubiquitin-activating enzyme. *Biochemistry,* **1994**, *33*(23), 7300-7308.

[23] Huang, H.D.; Lee, T.Y.; Tzeng, S.W.; Horng, J.T. KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. *Nucleic Acids Res.,* **2005**, *33,* W226-9.

[24] Pintar, A.; Carugo, O.; Pongor, S. Atom depth as a descriptor of the protein interior. *Biophys. J.,* **2003**, *84*(4), 2553-2561.

[25] Pintar, A.; Carugo, O.; Pongor, S. DPX: for the analysis of the protein core. *Bioinformatics,* **2003**, *19*(2), 313-314.

[26] Gáspári, Z.; Hudáky, I.; Czajlik, A.; Perczel, A. Is there an excuse for the non-conformist? Notes on the calculated energies, atom-atom contacts and natural abundance of the different conformers of alanine in proteins. *J. Mol. Struct. Theochem.,* **2004**, *675,* 141-148.

[27] Johnson, E.C.; Lazar, G.A.; Desjarlais, J.R.; Handel, T.M. Solution structure and dynamics of a designed hydrophobic core variant of ubiquitin. *Structure,* **1999**, *7*(8), 967-976.

[28] Desjarlais, J.R.; Handel, T.M. De novo design of the hydrophobic cores of proteins. *Protein Sci.,* **1995**, *4*(10), 2006-2018.

[29] Dosztanyi, Z.; Fiser, A.; Simon, I. Stabilization centers in proteins: identification, characterization and predictions. *J. Mol. Biol.,* **1997**, *272*(4), 597-612.

[30] Gromiha, M.M.; Pujadas, G.; Magyar, C.; Selvaraj, S.; Simon, I. Locating the stabilizing residues in (alpha/beta)8 barrel proteins based on hydrophobicity, long-range interactions, and sequence conservation. *Proteins,* **2004**, *55*(2), 316-329.

[31] Cornilescu, G.; Marquardt, J.L.; Ottiger, M.; Bax, A. Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. *J. Am. Chem. Soc.,* **1998**, *120*(27), 6836-6837.

Received: October 31, 2009 Revised: March 14, 2010 Accepted: July 28, 2010