Biotechnology and Bioprocess Engineering **2004**, 9: 414-418

Minimally Complex Problem Set for an *Ab Initio* Protein Structure Prediction Study

RyangGug Kim[1] and Cha-Yong Choi[1,2]*

[1] Interdisciplinary Program for Biochemical Engineering and Biotechnology, College of Engineering, Seoul National University, Seoul 151-742, Korea
[2] School of Chemical Engineering, College of Engineering, Seoul National University, Seoul 151-742, Korea

Abstract A "minimally complex problem set" for *ab initio* protein structure prediction has been proposed. As well as consisting of non-redundant and crystallographically determined high-resolution protein structures, without disulphide bonds, modified residues, unusual connectivities and heteromolecules, it is more importantly a collection of protein structures, with a high probability of being the same in the crystal form as in solution. To our knowledge, this is the first attempt at this kind of dataset. Considering the lattice constraint in crystals, and the possible flexibility in solution of crystallographically determined protein structures, our dataset is thought to be the safest starting points for an *ab initio* protein structure prediction study.

Keywords: *ab initio* protein structure prediction, crystal-packing effect, protein databank, dataset

Since Anfinsen [1] clarified that all the information required for a protein's folding was contained in its amino acid sequence, various approaches have been attempted to predict a protein's structure from its amino acid sequence alone, including comparative modeling, fold recognition and *ab initio* methods [2]. Among these, the *ab initio* approach is the only one that can be used when there is not even a remote homolog of a known structure for the target protein. Currently, more than half the proteins and two thirds of the known domains have no suitable template sequence for the prediction of their structure [3]. For these, the *ab initio* approach is the only one available for this end.

The *ab initio* prediction of the native structures of proteins is based on the thermodynamic hypothesis of protein folding [4,5], which states; "the native structure of a protein is the structure of its global energy minimum". Thus, the aim of current *ab initio* protein structure prediction is primarily to find the global energy minimum structure of the target protein, using some kind of energy or scoring function. The development of a scoring function needs the optimization of its parameters, using a database of known protein structures. In this respect, the quality of a scoring function may be affected by the protein structure dataset used in the optimization of its parameters.

Currently, the most extensive archive of protein structures is the protein databank (PDB) [6], which is composed of protein structures determined under diverse conditions. Besides the complications made by heteromolecules, disulphide bonds and modified residues, the crystal-packing effect can be a serious problem in an *ab initio* protein structure prediction. Current *ab initio* protein structure prediction methods use crystallographically determined static structures as their reference goal. However, long stretches or loops, which are stable in crystals, can be very mobile in solution, and the contact between the asymmetric units can affect the protein structure [7,8]. The protein structures containing these fragments, or determined under such a condition, would be different in crystal compared to solution. In fact, the structure of α-lactalbumin is different in crystal to in solution [9]. To our knowledge, no dataset has been complied with regard to the above criteria. Thus, an attempt was made to compile protein structures, which are simplest in their composition, and able to be used without worrying about the different conformations in the crystal form and when in solution. This dataset was termed the "minimally complex problem set" (MCPS). The MCPS is a collection of the simplest and safest protein structures for an *ab initio* protein structure prediction study.

The contacts between adjacent asymmetric units in a crystal were determined using the WHATIF server [10]. Heavy atom coordinates and a cutoff distance of 0.5 Å were used for the detection of the contacts between adjacent asymmetric units. The accessible surface area was calculated according to the method of Tsodikov *et al.* [11]. A probe radius of 1.4 Å and the van der Waals radii of Richards [12] were used. The radius of gyration of a protein was determined with only the positions of the alpha carbon atoms.

The process of MCPS compilation is summarized in

***Corresponding author**
Tel: +82-2-880-7071 Fax: +82-2-888-7295
e-mail: choicy@snu.ac.kr

Table 1. Construction of MCPS29

Stage	Selection criteria[a)]	Number of entries
1	Proteins from the initial PDB database	19,623
2	Crystal structure with a resolution less than 2.0 Å	7,371
3	Proteins of only a single chain	4,228
4	Disulphide bonds, modified residues, unusual connectivities, and heteromolecules are not contained	483
5	No duplicate or homologous entries	211
6	No disconnected amino acid sequences	154
7	No missing residues or atoms	64
8	No protruding N- and C-terminal parts making contacts with adjacent asymmetric units[b)]	38
9	No loops making extensive contacts with adjacent asymmetric units[c)]	32
10	No extensive contacts with adjacent asymmetric units relative to surface area[d)]	29

[a)] Selection criteria applied on the entries from the previous stage.
[b)] The entries containing the terminal residue stretches not making contact with the rest of the protein body (residues beyond 3 residues before or after the stretches) were removed.
[c)] The entries containing loops more than 6 residues long, with more than 60% contact with adjacent asymmetric units, were removed.
[d)] The entries, more than 60% of whose surface area makes contact with adjacent asymmetric units, were removed.

Table 1. The initial database was PDB release #103 (January 2003). The number of x-ray crystallographically determined structures of single-chain proteins, with a resolution of less than 2.0 Å, was 4,228. Of these, 2,001, 5,567 and 511 entries were found to contain disulphide bonds, heteromolecules other than water (*e.g.*, metal ions and other cofactors), and modified residues, respectively. Currently, heteromolecules can create a problem in an *ab initio* protein structure prediction, since it would be hard to determine which heteromolecules are required, and where they should be placed, to maintain the native structure of a given protein. The redundant entries, those with broken connectivities between amino acids, and those with unspecified positions of some of the constituting amino acids, were also removed. In sum, the number of remaining entries was 64.

The crystal-packing effect was then considered in three ways, namely the contacts of terminal residues with adjacent asymmetric units, those of loops and the ratio of the contact to surface areas. Since beta-turns are composed of 4 residues, we considered loops at least 6 residues long. If there exists a loop more than 6 residues long in crystal, and it does not make extensive contacts with adjacent asymmetric units, it would be highly possible that this loop structure could be maintained in solution, too. As for the quantitative criterion for the 'extensive' contact, we arbitrarily used 60% of the surface area. Thus, the entries containing loops which were more than 6 residues long and more than 60% of whose surface area made contact with adjacent asymmetric units were removed. The entries, more than 60% of whose surface area makes contact with adjacent asymmetric units, were also removed. After removing these entries, only 29 remained. The dataset composed of these 29 entries was termed; MCPS29. Some structural information on this MCPS29 is summarized in Table 2. The polypeptide length of the MCPS29 entries ranged from 61 (1I2T) to 679 (1SLL). There were some proteins in the same fold categories, but their sequence identities, as determined by FASTA [14], were less than 20%, with the exception of those for 1AAJ and 2PCY, which were both 22%.

Currently, the protein structures obtained by x-ray crystallography are used as the reference structures for *ab initio* protein structure prediction. However, in some cases, it has been reported that these structures are different from their solution structures [9]. Since all proteins operate in solution environment, the prediction of crystal-specific structures will be of less use in the end. Considering that there are not many protein structures determined by NMR, thus the selection of crystallographically determined structures identical to those in solution will be helpful in *ab initio* protein structure prediction study. We are proposing the MCPS29 in this respect.

As an example of the desirable structural properties of the MCPS29, the relationship between the polypeptide lengths and the radii of gyration in the MCPS29 was compared with that in the dataset obtained by applying the criteria of stages 6 and 7 on the stage 3 dataset in Table 1 (Fig. 1). Most entries were at the bottom of the polypeptide length-radius of gyration distribution, indicating that they were very compact. It shows that the MCPS29 does not have many extended and loose residues. The entries that significantly deviated from this trend were 1DVO, 1H6U, and 1JMW. The elongated structures of 1H6U [15] and 1JMW [16] are stabilized by their hydrophobic cores. Although 1DVO [17] has a long alpha helix at its N-terminal, only 4 residues at its terminus

Table 2. Structural properties of MCPS29

PDB code	Chain length	Class	Fold	Accessible surface area(Å²)		Radius of gyration(Å)
				Polar	Nonpolar	
1CEM	363	α	α/α toroid	6696.1	6148.2	18.75
2YGS	92	α	DEATH domain	2680.6	2528.1	11.60
1JMW	146	α	Four-helical up-and-down bundle	4041.4	4840.3	18.39
1I2T	61	α	PABP domain-like	1560.2	2398.0	10.55
1DVO	152	α	Repressor of bacterial conjugation FinO	4790.1	5639.9	23.13
1C5H	185	β	Concanavalin A-like lectins/glucanases	3976.1	3848.1	15.03
1DQ0	237	β	Concanavalin A-like lectins/glucanases	5335.5	5259.2	16.93
1SLL	679	β	Concanavalin A-like lectins/glucanases (81-276)[a]	12462.2	12090.6	27.54
		β	6-Bladed β-propeller(277-759)[a]			
1AAJ	105	β	Cupredoxin-like	2403.8	3106.3	12.33
2PCY	99	β	Cupredoxin-like	2555.0	2499.2	12.34
1AMM	174	β	γ-Crystallin-like	4737.8	3833.7	16.13
1KQX	134	β	Lipocalins	3762.3	3567.5	13.94
1BM8	99	α+β	DNA-binding domain of Mlu-1 binding protein MBP-1	2656.4	2832.7	12.07
1HKA	158	α+β	Ferredoxin-like	4023.6	4598.7	15.03
1AKI	129	α+β	Lysozyme-like	3501.5	3026.3	13.75
2LZM	164	α+β	Lysozyme-like	4213.8	4361.7	16.27
1EW4	106	α+β	N domain of copper amine oxidase-like	3275.4	2904.3	13.27
1AHC	246	α+β	Ribosomal inactivating proteins	5522.7	5710.7	17.54
1BV1	159	α+β	TBP-like	3998.2	4727.1	15.46
1SUR	215	α/β	Adenine nucleotide α hydrolase-like	5394.9	6218.6	17.56
1A8Q	274	α/β	α/β-Hydrolases	5356.3	5502.8	16.94
1AKZ	223	α/β	DNA glycosylase	4919.5	5533.2	16.70
1H6U	308	α/β	Leucine-rich repeat, LRR (right-handed β/α superhelix)	6174.0	7280.7	25.34
2PTH	193	α/β	Phosphorylase-hydrolase-like	4414.7	4941.7	15.28
1FSF	266	α/β	Phosphosugar isomerase	5427.0	6422.7	17.61
1SRV	145	α/β	The "swiveling" β/α/β domain	3709.3	3570.6	13.56
1CWY	500	α/β	TIM β/α-barrel	9383.6	11033.7	22.91
1UOK	558	α/β	TIM β/α-barrel(1-479)[a]	10987.5	10580.7	23.70
		β	α-Amylases, C-terminal β-sheet domain(480-558)[a]			
1SKF	262	α and β	β-Lactamase/D-ala carboxypeptidase	5081.0	5792.0	17.69

The structural class and fold of proteins were determined by SCOP 1.63 release [13].
[a] Each of these proteins contains two distinct folds. The numbers in brackets refer to the numbers of the starting and ending residues for each fold.

make contacts with the adjacent asymmetric units, suggesting that this elongated structure is not maintained by extensive contacts with the adjacent asymmetric units.

The criterion not considered in our compilation was the requirement for chaperones, which may be required for the folding of large or multimeric proteins [18]. When

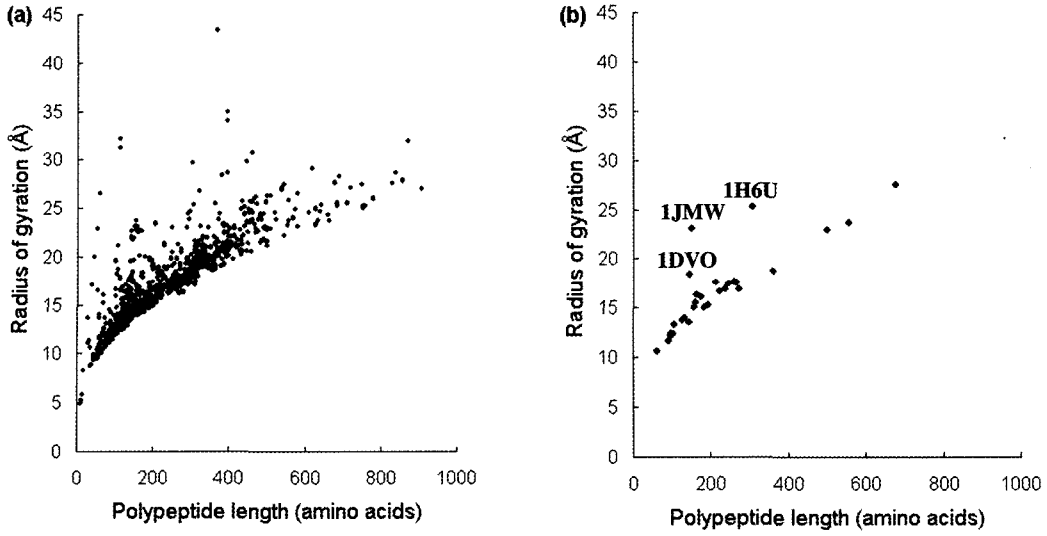

Fig. 1. The relationship between the polypeptide lengths and radii of gyration in MCPS29 (a) and the dataset obtained by applying the criteria of stages 6 and 7 on the stage 3 dataset (b).

the selection criteria of having only a single domain and a polypeptide length of less than 200 amino acids were applied to the MCPS29, only 17 entries remained (1AAJ, 1AKI, 1AMM, 1BM8, 1BV1, 1C5H, 1DVO, 1EW4, 1HKA, 1I2T, 1JMW, 1KQX, 1SRV, 2LZM, 2PCY, 2PTH and 2YGS). The dataset composed of these entries was termed; MCPS17. Compared with the size of the initial protein structure database (19,623 entries), those of MCPS29 and MCPS17 were surprisingly small.

Two datasets, MCPS29 and MCPS17, have been compiled, which are thought to be collections of the safest protein structures for an *ab initio* protein structure prediction. It is also highly possible that they might give good scoring functions for solution protein structures. This study is, to our knowledge, the first attempt to construct a protein structure dataset with regard to the difference in crystal and solution structures. Considering the vast complexity of the *ab initio* protein structure prediction field, it might be helpful to start with these safest dataset for an *ab initio* protein structure prediction.

REFERENCES

[1] Anfinsen, C. B., R. R. Redfield, W. L. Choate, J. Page, and W. R. Carroll (1954) Studies on the gross structure, cross-linkages, and terminal sequences in ribonuclease. *J. Biol. Chem.* 207: 201-210.

[2] Moult, J. (1999) Predicting protein three-dimensional structure. *Curr. Opin. Biotechnol.* 10: 583-588.

[3] Sánchez, R. and A. Sali (1998) Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc. Natl. Acad. Sci. USA* 95: 13597-13602.

[4] Anfinsen, C. B. (1973) Principles that govern the folding of protein chains. *Science* 181: 223-238.

[5] Gummadi, S. N. (2003) What is the role of thermodynamics on protein stability. *Biotechnol. Bioprocess Eng.* 8: 9-18.

[6] Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne (2000) The protein data bank. *Nucleic Acids Res.* 28: 235-242.

[7] Jacobson, M. P., R. A. Friesner, Z. Xiang, and B. Honig (2002) On the role of the crystal environment in determining protein side-chain conformations. *J. Mol. Biol.* 320: 597-608.

[8] Shimada, A., O. Nureki, M. Goto, S. Takahashi, and S. Yokoyama (2001) Structural and mutational studies of the recognition of the arginine tRNA-specific major identity element, A20, by arginyl-tRNA synthetase. *Proc. Nat. Acad Sci. USA* 98: 13537-13542.

[9] Urbanova, M., R. K. Dukor, P. Pancoska, V. P. Gupta, and T. A. Keiderling (1991) Comparison of alpha-lactalbumin and lysozyme using vibrational circular dichroism. Evidence for a difference in crystal and solution structures. *Biochemistry* 30: 10479-10485.

[10] Rodriguez, R., G. Chinea, N. Lopez, T. Pons, and G. Vriend (1998) Homology modeling, model and software evaluation: three related resources. *CABIOS* 14: 523-528.

[11] Tsodikov, O. V., M. T. Jr. Record, and Y. V. Sergeev (2002) A novel computer program for fast exact calculation of accessible and molecular surface areas and average surface curvature. *J. Comput. Chem.* 23: 600-609.

[12] Richards, F. M. (1977) Areas, volumes, packing and protein structure. *Annu. Rev. Biophys. Bioeng.* 6: 151-176.

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

[14] Pearson, W. R. and D. J Lipman. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA* 85: 2444-2448.

[15] Schubert, W. D., G. Gobel, M. Diepholz, A. Darji, D. Kloer, T. Hain, T. Chakraborty, J. Wehland, E. Domann, and D. W. Heinz (2001) Internalins from the human

pathogen Listeria monocytogenes combine three distinct folds into a contiguous internalin domain. *J. Mol. Biol.* 312: 783-794.
[16] Yu, E. W. and D. E. Koshland Jr. (2001) Propagating conformational changes over long (and short) distances in proteins. *Proc. Natl. Acad. Sci. USA* 98: 9517-9520.
[17] Ghetu, A. F., M. J. Gubbins, L. S. Frost, and J. N. Glover (2000) Crystal structure of the bacterial conjugation repressor *finO*. *Nat. Struct. Biol.* 7: 565-569.
[18] Deuerling, E., H. Patzelt, S. Vorderwulbecke, T. Rauch, G. Kramer, E. Schaffitzel, A. Mogk, A. Schulze-Specking, H. Langen, and B. Bukau (2003) Trigger factor and DnaK possess overlapping substrate pools and binding specificities. *Mol. Microbiol.* 47: 1317-1328.

[Received January 28, 2004; accepted September 28, 2004]