PDB-DO: Database of Disorder

Kim Lan Sim [1] klsim@ims.u-tokyo.ac.jp
Tomoyuki Uchida [2] uchida@cs.hiroshima-cu.ac.jp
Satoru Miyano [1] miyano@ims.u-tokyo.ac.jp

[1] Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan.
[2] Faculty of Information Sciences, Hiroshima City University, Hiroshima 731-3194, Japan.

1 Introduction

Disorder in protein structure and function have received very little attention even though the significance of these disordered regions from the functional aspects cannot be denied [2, 3, 5]. Protein Data Bank (PDB) which contains the largest collection of structural information of proteins is also the main source of disordered proteins. However, due to the lack of organised annotation and collective data on disordered regions from this database, the biological community at large is unaware of the importance of disorder. Hence, we propose here a database that will provide more accessibility to a systematic collection of disorder proteins from PDB.

2 PDB-DO

The database of disorder is located at the website [6]. Two searches are available: a) General search - based on selected keywords (disorder, gap, missing, poorly_ordered, flexible_linker, linker, flexible, unfolded, molten_globule and random_coil) (Fig. 1(b)). Source: PDB-DO. b) Specific Search - for detailed searches (Fig. 1(a)). Source: PDB (header text). HTML files are generated.

3 Results and Discussion

The PDB-DO is a "screened" database created based on selected keywords that denote disorder (or possible disorder) in proteins. Hence, this retrieval system reduces the task of biologists in collecting/extracting disorder entries from PDB. Besides, the system also provide a means for extracting data in a collective and global manner. A more rounded and refined search could be done in Specific Search. The search method employed in this system is a sequential text search system SIGMA [1] that enables very fine searching and editing of texts. Specific Search which realizes very fine AND/OR/NOT-search with keywords (any strings of symbols/codes) together with searching ability of patterns like KEYWORD1...KEYWORD2 (this finds a text which contains both KEYWORDS1 and KEYWORDS2 but KEYWORD1 occurs in the left of KEYWORD2 in the text.)

Future work includes categorising the entries generated from PDB-DO into a more systematic and detailed manner. This will be done by designing views using HypothesisCreator and finding common features (hydropathy, sequence complexity, etc.) in protein entries that share common "keywords". Our preliminary studies on knowledge extraction of disordered regions of proteins from PDB have proved the feasibility of this process [4]. We hope this retrieval system will help to bring about further awareness for biologists for future studies of disorder in proteins.

References

[1] Arikawa, S., Haraguchi, M., Inoue, H., Kawasaki, Y., Miyahara, T., Miyano, S., Oshima, K., Sakai, H., Shinohara, T., Shiraishi, S., Takeda, M., Takeya, S., Yamamoto, A., and Yuasa,

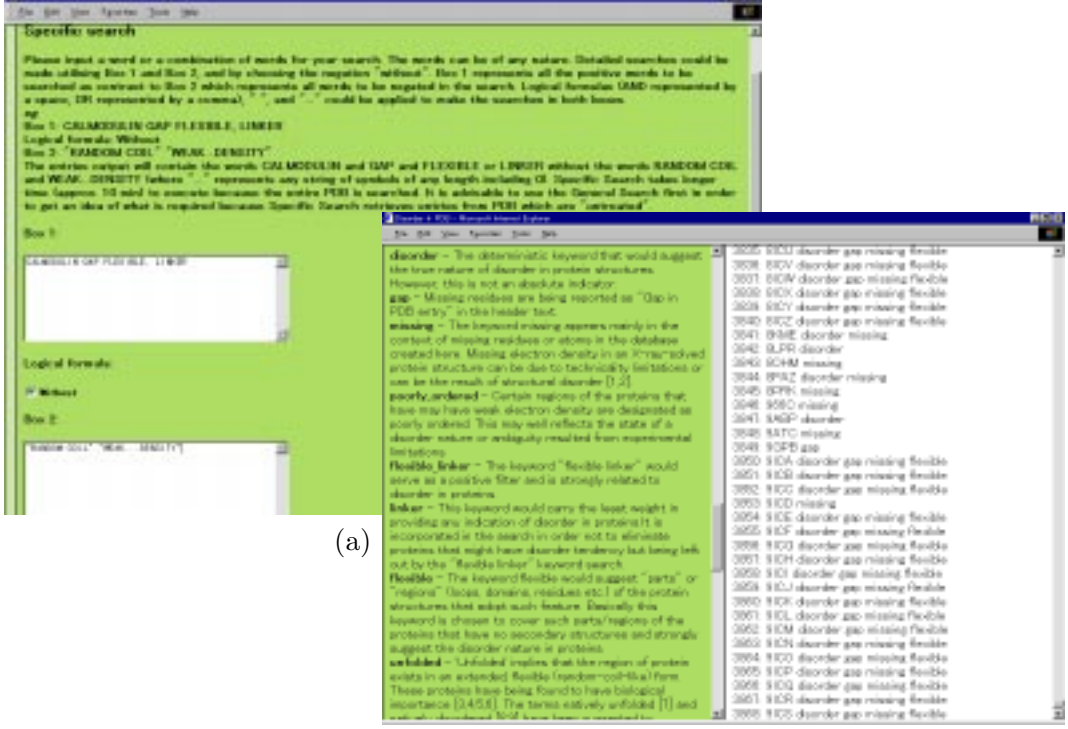

Figure 1: (a) Specific Search (b) An example of the output for General Search using the combination of all selected keywords. Reasons accompanying the selected keywords are given on the left webpage to aid users in giving weights in their searches. The different usages of the chosen keywords are also displayed in a separate webpage. It is meant to give users an idea of how the keywords are used in "disorder" context. Entries with the most keyword hits are likely to reflect the highest possibility of disorder.

H., The text database management system SIGMA: an improvement of the main engine, *Proc. Berliner Informatik-Tage*, 72–81, 1989.

[2] Daughdrill, G.W., Chadsey, M.S., Karlinsey, J.E., Hughes, K.T. and Dahlquist, F.W., The C-terminal half of the anti-sigma factor, FlgM, becomes structured when bound to its target, *Nature Struc. Biol.*, 4:285–291, 1997.

[3] Dunker, A.K., Garner, E., Guilliot, S., Romero, P., Albrecht, K., Hart, J., Obradovic, Z., Kissinger, C., and Villafranca, J.E., Protein disorder and the evolution of molecular recognition: theory, predictions and observations, *Proc. Pacific Symposium on Biocomputing 1998*, World Scientific, 3:471–482, 1998.

[4] Maruyama, O., Uchida, T., Sim, K.L., and Miyano, S., *Lecture Notes in Artificial Intelligence (Proc. First International Conference on Discovery Science)*, 1532:105–116, Springer-Verlag, 1998.

[5] Weinreb, P.H., Zhen, W., Poon, A.W., Conway, K.A. and Lansbury, Jr. P.T., NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded, *Biochemistry*, 35(43):13709–13715, 1996.

[6] http://bonsai.ims.u-tokyo.ac.jp/~klsim/GIW99_disorder.html