# Annotation of PDB with respect to "Disordered Regions" in Proteins

**Meeta Rani** [1] meeta@bic.nus.edu.sg
**Pedro Romero** [2] promero@eecs.wsu.edu
**Zoran Obradovic** [2] zoran@eecs.wsu.edu
**A. Keith Dunker** [3] dunker@mail.wsu.edu

[1] BioInformatics Centre, National University of Singapore, 10 Kent Ridge Crescent, Singapore 11926099
[2] Department of Electrical Engineering and Computer Science, Washington State University, Pullman, WA, 99164, USA
[3] Department of Biochemistry and Biophysics, Washington State University, Pullman, WA, 99164, USA

## 1 Introduction

Protein function has conventionally been associated with three-dimensional structure, with the understanding that folding is a prerequisite to function. On the contrary, there exist proteins containing regions that do not fold as expected but are required for important biological functions. We call these "disordered" regions.

Disorder provides the basis for numerous functions, but an especially interesting one is the involvement of disorder in molecular recognition. This appears contrary to the long-held views of complementary surface fitting for binding, but such disordered regions typically become ordered as the proteins associate with their cognate molecules thereby leading to surface complementarity in the complex. Well-characterized examples include enzyme-substrate, receptor-ligand, protein-peptide, protein-protein, protein-RNA and protein-DNA complexes. It has been suggested that disorder-order transitions allow biologically advantageous combination of high specificity coupled with low affinity. If disordered regions are so involved in function then they are likely to have distinctive amino acid sequences.

Disordered regions in proteins can be random coil-like, molten globule-like or somewhere in between. Such regions can be identified by hypersensitivity to protease digestion, absence of coordinates in the x-ray data and characterization by CD or NMR spectroscopy. Using these methods, hundreds of examples of disordered regions are now known.

We have initiated work on the sequence-disorder relationships [1]. The single greatest difficulty in our studies is the lack of organized information on disordered regions. For example, although the Protein Data Bank probably contains the largest collection of information regarding disordered proteins, this information is very difficult to retrieve as the disordered regions are not reported or annotated in any uniform way [2]. Even worse is that disordered regions are sometimes obscured by the use of model building to add the missing coordinates.

Given the growing realization of the importance of disordered regions, we feel it is very timely and useful to annotate and classify the proteins in the PDB with specific reference to disorder. Since disordered regions are unobservable in the x-ray data, this property can be used as a criterion for marking a PDB entry as containing a disordered region and then for further classification based on this information. However missing coordinates in an X-ray determined structure could arise from multiple causes, not all of which imply flexible disorder (eg., random coil-like or molten globule-like structures). Therefore annotation of theproteins of the PDB to include information on disorder should be regarded as a first step to which information from other sources can be added eventually.

## 2 Identification and analysis of disordered regions in the PDB

Our study involves three steps: identification, annotation and classification of the disordered regions in the PDB. These 3 steps are described briefly as follows:

*A: Identification of disordered regions*

The coordinates of the atoms are usually not reported in the PDB files. Thus such PDB entries show fewer amino acids than contained in the actual sequence. To find all the disordered regions in the PDB, the number of amino acids in the structure files are compared with the actual number of amino acids in them. All proteins for which there are discrepancies in these two data are saved. Next, the amino acids in the co-ordinate list are compared one-by-one to amino acids in the sequence. Sequence discrepancies are flagged for manual investigation and amino acids that are missing from the coordinate lists are identified to constitute disordered regions in the proteins.

*B: Annotation of the disordered regions*

A new keyword, say, DS (short for "disorder status") will be added to each entry and the following information will be inserted in each record. DS: (disorder status): a disordered region is PRESENT or ABSENT. If DS = PRESENT then a) Position of DR in the protein: position x to position y. b) Vicinity of DR: C-terminus/N-terminus/ internal to sequence c) Class: based on length of DR (see below). d) Percentage of amino acids in DR: x% of length. e) Function: confirmed or putative description of the function. f) Comments: Important information regarding further characterization of the given DR, by NMR or such as hypersensitivity to protease action may be added.

*C: Classification of the disordered regions*

Our initial classification of the disordered regions will be based on their lengths. The rationale is that different lengths of disordered regions will have different potential capabilities. Our tentative grouping will be: a) Very short disordered regions, 1-8 AA. b) Short disordered regions, 9-19 AA. c) Medium disordered regions, 20-39 AA. d) Long disordered regions 40-59 AA e) Very long disordered regions $\geq 60$ AA.

The annotation and the length-based classification of the disordered regions will be included in the records of the PDB entries and the new database will be called Pdb_disorder_Plus. This annotated database will provide an organized collection of information on disordered regions in proteins and is expected to encourage further investigation of disordered regions and their putative functions. On completion, (soon expected to be), Pdb_disorder_Plus will be jointly shared and maintained by the BioInformatics Centre, National Univ. of Singapore at their home-page (http://www.bic.nus.edu.sg/),and Department of Biochemistry and Biophysics, WSU, Pullman, USA, at their "Protein Disorder homepage" (http://www.disorder.chem.wsu.edu/).

## References


[1] Romero P., Obradovic Z., Kissinger C., Villafranca J. E., and Dunker A. K., Identifying Disordered Regions in Proteins from Amino Acid Sequences, *Proceedings of the International Conference on Neural Networks (ICNN'97)*, at Houston, Texas, USA, 90–95, 1997.

[2] Dunker, A.K., Obradovic, Z., Romero, P., Kissinger, C., and Villafranca,J. E., On the Importance of Being Disordered, *PDB Newsletter*, 81:3–5, 1997.