ProDDO: a database of disordered proteins from the Protein Data Bank (PDB)

Kim Lan Sim1,∗, Tomoyuki Uchida2 and Satoru Miyano1

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

Received on March 2, 2000; revised on October 20, 2000; accepted on November 1, 2000

ABSTRACT

Summary: ProDDO represents a ‘pre-screened’ database that denotes disorder (or possible disorder) in proteins from the PDB.

Availability: ProDDO is available at http://bonsai.ims.u-tokyo.ac.jp/~klsim/database.html

Contact: klsim@plato.gws.uky.edu

Disordered regions in proteins are regions which do not conform to the usual secondary structures. The terms natively unfolded (Weinreb et al., 1996), natively disordered (Dunker et al., 1998) and intrinsically unstructured have been coined to describe such phenomenon. The functional significance of disordered regions can be found in numerous important biological functions such as DNA-induced folding, transcriptional activation, translational processes, cell regulation and amyloid formation in prion protein (Wright and Dyson, 1999; Fletcher et al., 1998; Donne et al., 1997). Despite its significance, there has been a lack of organised annotations of disordered regions in proteins in a collective and global manner apart from the work of Romero et al. (1997). Our proposed ProDDO provides accessibility to a systematic collection of disordered proteins from the PDB in a global manner.

The creation of ProDDO was based on through biological screening of significant keywords (Table 1) that denote disorder in the real context. The system itself gives users (biologists or bioinformatists) flexibility to create their own disordered database from the already ‘pre-screened’ database. Users can select the keywords to input to generate the required database. Each keyword is linked to a separate database which describes the different usages of that particular keyword in the entire PDB giving suggestions in the context of ‘disorder’.

ProDDO is written in Python which calls C and C++ programs. The search engine and database management system of ProDDO employs a Unix version (written in C) of a general purpose text database management system SIGMA (Arikawa et al., 1990) that exploits very fast and efficient one-way sequential processing of text files. ProDDO consists of General Search (disordered proteins) and Specific Search (any protein searches). Specific Search performs a very fine AND/OR/NOT-search with keyword patterns (any strings of symbols/codes) together with searching ability of patterns like KEYWORD1...KEYWORD2. This fine-sequential search generates more hits as compared to PDB that utilises inverted files. The output in both searches are displayed interactively in html files.

PDB which is the main source of disordered proteins is highly biased, due to the fact that only proteins which crystallise well are being represented (Gerstein, 1998). Moreover, variants or homologous structures create redundancy and biasness in the amino-acid complexity deposited in it. In ProDDO, retrieval based on the selected keywords overcomes these problems. Primarily, the search does not take into account the amino-acid complexity, and regions in proteins that do not crystallise well but the missing coordinates are noted in the header text can be retrieved. A protein having homologous structures that exist in multiple forms but having disorder/order discrepancy can also be filtered out. It is very difficult to extract disordered data from >10,000 entries in PDB. ProDDO contains 3879 entries but the actual disordered database created can be as small as 5–10 entries depending on the threshold of the number of keyword hits imposed by the user.

Example:
The output from General Search:
1NLR disorder missing flexible linker linker flexible
1ZQG disorder gap missing flexible
9ICY gap

In the example above, 1NLR (5 hits) is a better candidate than 1ZQG (4 hits) or 9ICY (1 hit). Several criteria were imposed on the selections of protein structures which are reported as disorder in...
developing ProDDO. Keywords that do not describe disorder in its real context were eliminated.

Example:

2NGR CDC42—GAP

The GAP keyword in the above example is a gene. Side chain and 1–2 atoms disorder are not considered to represent true disorder. Dynamic and static disorder, including N- and C-terminal disorder are taken as representatives of disorder. Distinctions on end-terminal disorder can only be made experimentally as to whether they are artifacts (due to technical limitations) or true disorder as in the case of PrP\textsuperscript{C} N-terminal octarepeat which is a disordered region (Donne \textit{et al.}, 1997).

Our preliminary studies on knowledge extraction of disordered regions of proteins from PDB have proven the feasibility of this process (Maruyama \textit{et al.}, 1999). We hope ProDDO will help to bring about further awareness in biologists for future studies of disorder in proteins.

ACKNOWLEDGEMENTS

Kim Lan Sim wishes to thank the Japan Society for the Promotion of Science (JSPS) for their generous support.

REFERENCES


### Table 1. The ten selected keywords and reasons for their selections

<table>
<thead>
<tr>
<th>Keywords</th>
<th>Reasons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disorder</td>
<td>The deterministic keyword that would suggest the true nature of disorder in protein structures. Dynamic and static disorder are often reported in X-ray diffraction data (Bennett and Huber, 1984). Both types represent true disorder. However, this is not an absolute indicator.</td>
</tr>
<tr>
<td>Gap</td>
<td>Missing residues are being reported as ‘Gap in PDB entry’ in the header text.</td>
</tr>
<tr>
<td>Poorly ordered</td>
<td>Certain regions of the proteins that may have weak electron density are designated as poorly ordered. This may well reflect the state of a disorder nature or ambiguity resulting from experimental limitations.</td>
</tr>
<tr>
<td>Flexible_linker</td>
<td>The keyword ‘flexible linker’ would serve as a positive filter and is strongly related to disorder in proteins.</td>
</tr>
<tr>
<td>Linker</td>
<td>This keyword would carry the least weight in providing any indication of disorder in proteins. It is incorporated in the search in order not to eliminate proteins that might have disorder tendency but being left out by the ‘flexible linker’ keyword search.</td>
</tr>
<tr>
<td>Flexible</td>
<td>The keyword flexible would suggest ‘parts’ or ‘regions’ (loops, domains, residues, etc.) of the protein structures that adopt such features. Basically this keyword is chosen to cover such parts/regions of the proteins that have no secondary structures and strongly suggest the disorder nature in proteins.</td>
</tr>
<tr>
<td>Unfolded</td>
<td>‘Unfolded’ implies that the region of proteins exists in an extended, flexible (random-coil-like) form. These proteins have been found to have biological importance (Huth \textit{et al.}, 1997). The terms natively unfolded (Weinreb \textit{et al.}, 1996) and natively disordered (Dunker \textit{et al.}, 1998) have been suggested to represent disorder in proteins.</td>
</tr>
<tr>
<td>Molten_globule</td>
<td>A sequence that does not fold into a single unique 3-D structure under physiological conditions; can take the form of partially-folded, like a molten globule (Dolgikh \textit{et al.}, 1981).</td>
</tr>
<tr>
<td>Random_coil</td>
<td>One of the forms that a natively unfolded sequence adopts. The disordered ensemble of structures can involve equilibria between random-coil-like and molten-globule-like forms.</td>
</tr>
</tbody>
</table>