The Protein Coil Library: A Structural Database of Nonhelix, Nonstrand Fragments Derived from the PDB

Nicholas C. Fitzkee, Patrick J. Fleming, and George D. Rose
T. C. Jenkins Department of Biophysics, Johns Hopkins Univeristy, Baltimore, Maryland

ABSTRACT **Approximately half the structure of folded proteins is either α-helix or β-strand. We have developed a convenient repository of all remaining structure after these two regular secondary structure elements are removed. The Protein Coil Library (http://roselab.jhu.edu/coil/) allows rapid and comprehensive access to non-α-helix and non-β-strand fragments contained in the Protein Data Bank (PDB). The library contains both sequence and structure information together with calculated torsion angles for both the backbone and side chains. Several search options are implemented, including a query function that uses output from popular PDB-culling servers directly. Additionally, several popular searches are stored and updated for immediate access. The library is a useful tool for exploring conformational propensities, turn motifs, and a recent model of the unfolded state. Proteins 2005;58:852–854.** © 2005 Wiley-Liss, Inc.

Key words: protein structure; protein conformation; secondary structure; random coil; statistical coil

INTRODUCTION

The structures of folded proteins are inherently complex, and many cognitive schemes have been developed to simplify and organize protein substructure. Cartoon illustrations that reduce α-helices and β-strands to visual icons[1] have been especially useful tools because approximately half of any given folded protein adopts either or both of these two regular secondary structure motifs. Here, we focus on the other half of the protein, i.e., the "coil" regions.

The intriguing hypothesis that coil regions are apt models for the unfolded state of proteins has motivated several important studies. Swindells et al. distinguished between α-helices, β-strands, polyproline-II helices and coil (everything else) when calculating conformational propensities for amino acids.[2] Serrano compared the ϕ torsion angle propensities found in the coil conformation to NMR measurements of the unfolded state,[3] an approach that has been pursued in very recent work.[4,5] On the whole, however, comparatively few investigators have capitalized on the wealth of structural information stored in coil fragments.

The Protein Coil Library (PCL) is designed to address this issue. It classifies protein structure using a torsion-angle based standard and stores non-helix, non-strand fragments in an online database. The library includes molecular coordinates, dihedral angles, and sequence information for each fragment, and users can browse this information using a convenient web interface. Data can also be accessed via FTP. Versatile search tools are provided via a queued system, and the output from several online PDB-culling servers can be used to select the list of proteins to be included in a search. Additionally, the library provides basic utility programs to assist users in analyzing their search results.

Implementation

Secondary structure classification

The method used to classify secondary structure in the PCL, similar to that described by Srinivasan and Rose,[6] tiles Ramachandran dihedral space[7] into a course-grained 30° × 30° ϕ, ψ-grid. We refer to these grid squares as *mesostates*; each is assigned a unique identifier. Any protein backbone conformation can be approximated by its linear sequence of mesostate identifiers, and regular expressions of mesostate sequences can be used to define α-helices, β-strands, and turns. Hydrogen bonds are not included in our method, but, nevertheless, the results are in close agreement with those of other secondary structure classification programs (e.g. DSSP[8]) that do utilize hydrogen bonds. Mesostate bins are illustrated in Figure 1, overlaid on to a contour plot of Ramachandran dihedral angles calculated by Hovmöller et al.[9] The regular expressions used to define secondary structures (α-helix, β-strand, polyproline-II helix, turns, and coil) are given on the PCL web page and in the supplemental material.

Coil fragment excision

Using the secondary structure classification algorithm described above, nonhelix and nonstrand fragments were extracted from the Protein Data Bank.[10] Each fragment was inspected for chain breaks. Residues lacking any backbone atom (N, CA, C, or O) and single-residue fragments were excluded from the library. As a result, all fragments in the PCL are continuous and include at least

---

The Supplementary Materials referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/index.html

*Correspondence to: George D. Rose, T. C. Jenkins Department of Biophysics, The Johns Hopkins Univeristy, 3400 N. Charles Street, Baltimore, MD 21218-2608. E-mail: grose@jhu.edu

Received 19 October 2004; Accepted 22 October 2004

Published online 18 January 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20394

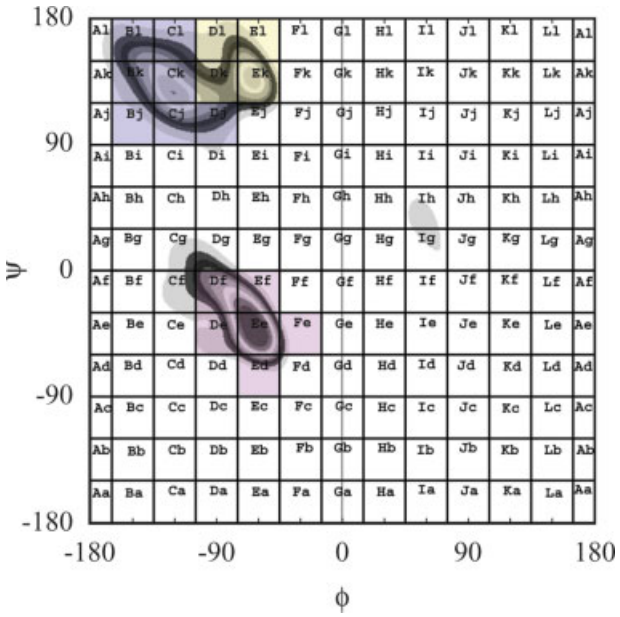

Fig. 1. A contour plot of Ramachandran dihedral space[9] overlaid with mesostate tile definitions. Mesostate identifiers are two character strings (e.g. Ae): the first character indicates a region along the ϕ axis and the second character indicates a region along the psgr; axis. Mesostates used to identify β-strands are shaded blue, those used for identifying α-helices are shaded pink, and those used for identifying polyproline-II are yellow. Mesostates Dl and Dk, while normally used to identify polyproline-II, will be classified as β-strand if adjoining residues are also β-strand. For more details, please see the web site.

two residues. Where possible, up to two flanking residues at both the N- and C-termini were also extracted to provide the context of the fragment. The resulting coordinates were stored in standard PDB format.

Torsion angle calculations

Accompanying every fragment is a data file that includes the sequence of the fragment along with the ϕ, ψ, ω, τ, and χ_n torsion angle values for each residue, according to the IUPAC-IUB standard.[11] The file also includes the per-residue mesostate identifiers and secondary structure classifications. The file format is designed to ease high-capacity analysis and is described in detail on the PCL website.

File naming and organization

Data from the coil library are stored in a collection of compressed text files that can be accessed via the web or anonymous FTP. Filenames reflect the origin and details of each fragment: the PDB identifier, chain, fragment length and start residue are all reported within each file name. All files are organized hierarchically by PDB ID and fragment length to minimize strain on the server file system. File naming conventions are described in detail on the website.

Interface and Usage

The Protein Coil Library can be accessed at http://roselab.jhu.edu/coil/. For simple searches, single chains from a PDB identifier may be browsed interactively. For each chain, coil fragments are listed and ranked according to size. The browsing functionality allows the user to download molecular coordinates directly or to view dihedral angle and secondary structure data in an HTML document. A cross-reference link to the Protein Data Bank site is associated with each coil fragment.

For more complex queries, a batch search form is provided that allows users to specify fragment sizes in addition to PDB and chain identifiers. In addition to a simple text file containing PDB ID's, PDBSelect,[12] and PISCES[13] formatted lists may be uploaded that specify which chains to include in the search. Using a PDBSelect or PISCES list allows the user to filter fragments based on sequence identity, resolution, and refinement quality (R-value). Once submitted, batch searches are queued, and when the results have been calculated, the user is notified that the search results are available on the server. Results are returned as a list of fragments stored on the server as well as a compressed archive of the dihedral angle data for all matched fragments. Coordinates for search results must be downloaded separately or extracted from a local copy of the PDB using one of the included utilities. Search results are removed from the server after two weeks.

Given the popularity of PISCES, two lists are generated automatically to ease resource consumption. The first list contains fragments extracted from PDB entries with a 90% sequence identity cutoff, a resolution of 2.0 Å or better, and an R-value of 25% or better. The second list contains fragments with a 20% sequence identity cutoff, a resolution of 1.5 Å or better, and an R-value of 25% or better. The results from these searches are always available as precompiled lists, and as demand arises other searches can be scheduled automatically as well. While the coil library itself is updated nightly from the PDB, these lists are only updated weekly, in coordination with the distribution of new PISCES lists.

Finally, a repository of analysis tools is provided on the website. In addition to a utility that will extract structural coordinates given a dihedral angle file, a tool is provided that can catalog the number of times different structural motifs appear in a dataset. As additional tools are implemented or contributed, they will be posted at this location.

Statistics

There are presently 784,257 coil fragments contained in the PCL, representing 55,111 chains in 25,392 unique PDB identifiers. The culled list containing fragments having less than 90% sequence identity cutoff, resolution of 2.0 Å or better, and an R-value better than 25% currently has 57,402 fragments representing 3,959 chains in 3,652 unique PDB identifiers. The distributions of fragment sizes for both lists are markedly skewed toward short fragments (Fig. 2). This is not surprising in light of hydrogen bonding considerations: α-helix and β-strand are the only regular structures that can satisfy hydrogen bonds for long chain segments, and the PCL lacks these structures. However, hydrogen-bonded structures are also abundant in short chain

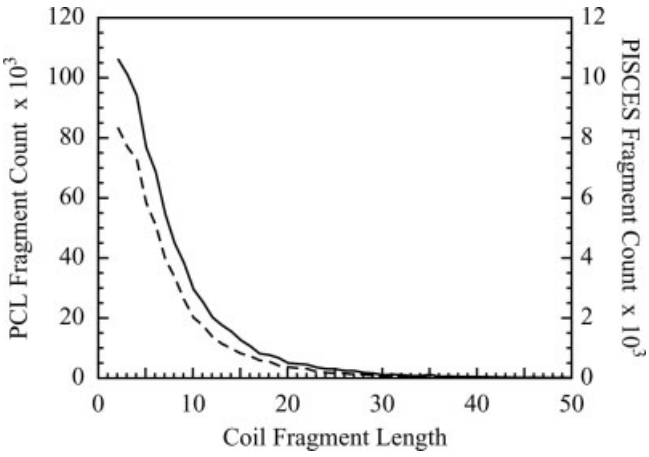

Fig. 2. A plot comparing the coil fragment length in the PCL to the number of times that length occurs. The solid line (left axis) shows the distribution for the entire coil library, and the dashed line (right axis) shows the distribution for a culled list of chains (90% sequence identity cutoff, 2.0 Å resolution or better, and 25% or better R-value). Both graphs indicate a decline in frequency as fragment size increases.

fragments. Indeed, using the least stringent hydrogen bond definitions outlined by Kortemme et al.,[14] approximately 40% of the residues in the PCL are involved in an i to $i+3$ hydrogen-bonded turn.

ACKNOWLEDGMENTS

We thank Nick Panasik, Timothy Street, and Haipeng Gong for their assistance in critiquing and debugging this work. Financial support from the Mathers Foundation is gratefully acknowledged.

REFERENCES

1. Kraulis PJ. Molscript—a program to produce both detailed and schematic plots of protein structures. J Appl Crystallogr 1991;24: 946–950.
2. Swindells MB, MacArthur MW, Thornton JM. Intrinsic phi, psi propensities of amino acids, derived from the coil regions of known structures. Nat Struct Biol 1995;2:596–603.
3. Serrano L. Comparison between the phi distribution of the amino acids in the protein database and NMR data indicates that amino acids have various phi propensities in the random coil conformation. J Mol Biol 1995;254:322–333.
4. Avbelj F, Baldwin RL. Origin of the neighboring residue effect on peptide backbone conformation. Proc Natl Acad Sci USA 2004;101: 10967–10972.
5. Fleming PJ, Fitzkee NC, Mezei M, Srinivasan R, Rose GD. A novel method reveals that solvent water favors polyproline II over beta-strand conformation in peptides and unfolded proteins: Conditional Hydrophobic Accessible Surface Area (CHASA). Protein Sci 2005;14:111–118.
6. Srinivasan R, Rose GD. A physical basis for protein secondary structure. Proc Natl Acad Sci USA 1999;96:14258–14263.
7. Ramachandran GN, Sasisekharan V. Conformation of polypeptides and proteins. Adv Prot Chem 1968;23:283–438.
8. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
9. Hovmöller S, Zhou T, Ohlson T. Conformations of amino acids in proteins. Acta Crystallogr D Biol Crystallogr 2002;58(Pt 5):768–776.
10. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
11. IUPAC-IUB. IUPAC-IUB Commission on Biochemical Nomenclature. Abbreviations and symbols for the description of the conformation of polypeptide chains. J Mol Biol 1970;52:1–17.
12. Hobohm U, Sander C. Enlarged representative set of protein structures. Protein Sci 1994;3:522–524.
13. Wang G, Dunbrack RL, Jr. PISCES: a protein sequence culling server. Bioinformatics 2003;19:1589–1591.
14. Kortemme T, Morozov AV, Baker D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. J Mol Biol 2003;326:1239–1259.