D320–D325 *Nucleic Acids Research, 2010, Vol. 38, Database issue* *Published online 11 November 2009*
doi:10.1093/nar/gkp1013

Protein Geometry Database: a flexible engine to explore backbone conformations and their relationships to covalent geometry

Donald S. Berkholz[1], Peter B. Krenesky[2], John R. Davidson[2] and P. Andrew Karplus[1,*]

[1]Department of Biochemistry and Biophysics, Oregon State University, 2011 ALS and [2]Open Source Lab, Oregon State University, B211 Kerr Admin, Corvallis OR 97331, USA

Received August 14, 2009; Revised October 14, 2009; Accepted October 19, 2009

ABSTRACT

The backbone bond lengths, bond angles, and planarity of a protein are influenced by the backbone conformation (φ,ψ), but no tool exists to explore these relationships, leaving this area as a reservoir of untapped information about protein structure and function. The Protein Geometry Database (PGD) enables biologists to easily and flexibly query information about the conformation alone, the backbone geometry alone, and the relationships between them. The capabilities the PGD provides are valuable for assessing the uniqueness of observed conformational or geometric features in protein structure as well as discovering novel features and principles of protein structure. The PGD server is available at http://pgd.science.oregonstate.edu/ and the data and code underlying it are freely available to use and extend.

INTRODUCTION

With the explosion in the number of atomic-resolution protein structures in the past decade, the possibility to determine accurate details of protein geometry from proteins themselves rather than from small-molecule peptides has become a reality. The importance of bond angles, bond lengths, and peptide planarity in validating structures as well as discovering real and functionally important deviations from standard geometry has become increasingly apparent (1–7). To our knowledge, no database has existed until now to search peptide geometry either on a large scale to discover trends or on an individual basis to explore unusual features. Unusual features that are significant often pass unrecognized even by the structural biologists who solved the structure (Figure 1).

The protein backbone confrmation is defined primarily by the dihedral angles φ and ψ together with whether the peptide bond is in the *trans* (ω near 180°) or *cis* (ω near 0°) conformation. The protein backbone geometry, on the other hand, is defined by the bond angles and lengths and deviations of the peptide from planarity. It has been shown that the average values for backbone geometry vary in a conformation-dependent manner, reflecting an intimate relationship between conformation and geometry (7). The prevalent misconception that bond angles and lengths are static has been caused in part by the lack of any straightforward way to examine their dependence on local conformation. The Protein Geometry Database (PGD) is a unique resource that now makes it possible for biologists to explore peptide geometry, peptide conformation, and the ties between them. Other databases exist to allow searching conformation alone [e.g. SPASM (8), Fragment Finder (9), Protein Segment Finder (10), Conformational Angles Database (11), PDBeMotif (12)], but even in this arena, the PGD offers a unique combination of convenience and flexibility.

IMPLEMENTATION

The PGD contains derived data for a complete, representative data set of protein structures that are relevant to discovering reliable instances of conformations and peptide geometries. This allows users to set thresholds specific to their queries at search time without being unduly limited by the cutoffs chosen during database creation. To ensure that the PGD data are representative of conformational and geometric space rather than being biased by multiple highly similar structures, the PGD contains data derived from a nonredundant set of proteins. As is common, the nonredundancy is defined by the maximum allowed sequence identity between any pair of proteins in the data set. Two thresholds of 25 and 90% are available in the PGD. The nonredundant set is taken from PISCES (13). Because different resolution ranges are suitable for different queries, the PGD

*To whom correspondence should be addressed. Tel: +1 541 737 3200; Fax: +1 541 737 0481; Email: karplusp@science.oregonstate.edu maximizes structural data by using all data included in the PISCES data sets, corresponding to crystal structures determined at 3.0 Å resolution or better with no cutoff for the crystallographic R-factor. Although the lower-resolution structures in this sample do have lower accuracy, users can easily exclude them using search parameters.

The PGD contains data on per-chain and per-residue levels. For each chain, stored parameters include the PDB code, the chain ID, the sequence-identity threshold, the resolution, and the crystallographic R-factor. The sequence-identity threshold, resolution, and R-factor are all useful in parameters to define the independence and quality of the data searched. For each residue, stored parameters include a mapping back to the chain and protein, the residue number, the torsion angles φ, ψ, ω and χ_1, the improper dihedral ζ [describing the chirality of the C_α (14)], all seven backbone bond angles, all five backbone bond lengths, the DSSP-defined (15) secondary-structure type, and three B-factors: the mainchain average, the sidechain average, and the C_γ atom. The B-factors are useful as cutoffs to exclude residues with poorly defined conformation and geometry [e.g. (7,16)].

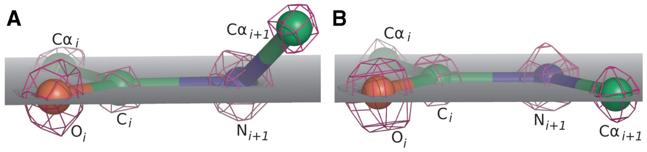


Figure 1. An active-site peptide geometry feature discovered using the PGD. (**A**) Shown is the peptide bond between residues His306 and Asn307 in the 0.90 Å resolution structure of Cu-nitrite reductase [PDB code 2bw4 (21)]. $2F_o$–F_c electron density is shown at 6.5 ρ_{rms} (violet mesh), and a plane is shown for reference. (**B**) Same as A but showing the peptide bond between residues Ala291 and Phe292. For standard planar peptides such as the one in panel B, all five atoms shown lie in a plane. In contrast, in panel A, the Cα atom of residue $i+1$ is highly deviant from the plane defined by the $C\alpha_i$, C, O, and N atoms. The electron density indicates that the atoms are all reliably positioned. This peptide bond is 37° from planar (ω = 143°). His306 ligates the Cu^{2+}, so this is an important structure-function feature that was overlooked. This example is not unique; residues with unrecognized yet real and important deviations in peptide geometry are relatively common.

The PGD uses the Python-based Django framework for both populating and searching a MySQL database. Using Django allows us to follow the DRY principle ('Don't Repeat Yourself') by only having one description of the database format. This reduces the difficulty of changes, increases the clarity of code, and avoids potential conflicts between multiple descriptions. A single change can transform the database schema for all applications that use it. The database is populated by interfacing with a tool written with BioPython (17) to calculate PDB-derived information. The tool, Splicer, splices derived data from the PDB files together into all possible consecutive segments from 1 to 10 residues long. This approach speeds searching because segments do not need to be constructed during every search.

A single run of Splicer to populate the PGD can take ~16 h on a current single-processor compute node, so we constructed a new Python framework for distributed, parallel data processing called Pydra <http://pydra-project.osuosl.org/>. Using Pydra, a parallel Splicer run across 20 CPUs on four nodes takes ~1 h, providing the nearly linear speedup expected for this type of coarse-grained parallelism.

The current version of the PGD contains ~3.8 million residues from nearly 16 000 protein chains, with all amino acids and secondary-structure types being well-represented (Figure 2). The PGD content will be updated on a quarterly basis or better.

SEARCHING AND ANALYZING RESULTS

The search page

The PGD has a professionally designed, user-friendly yet flexible graphical interface for mining protein conformational and geometric space. Upon proceeding beyond the introductory entry page, users encounter the search page (Figure 3). On this page, users define all parts

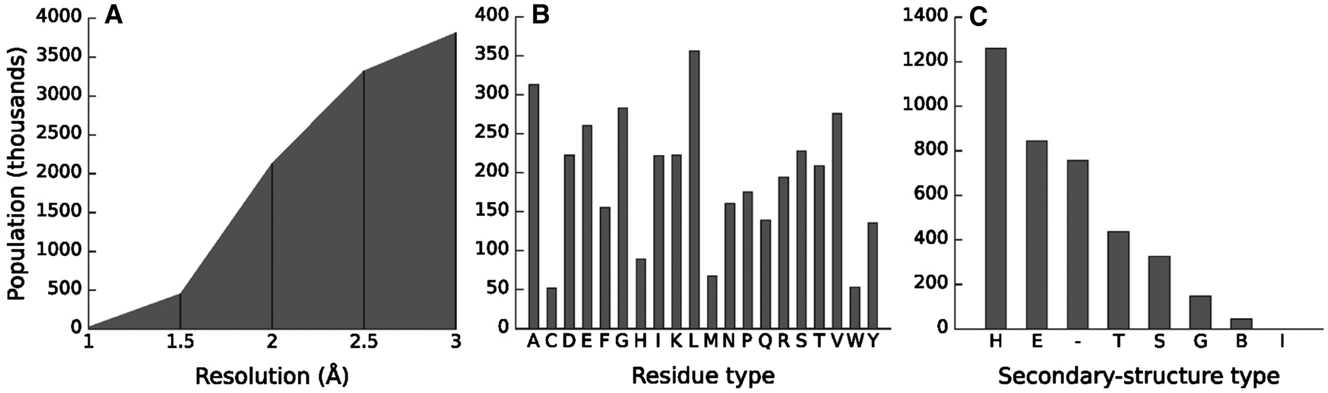
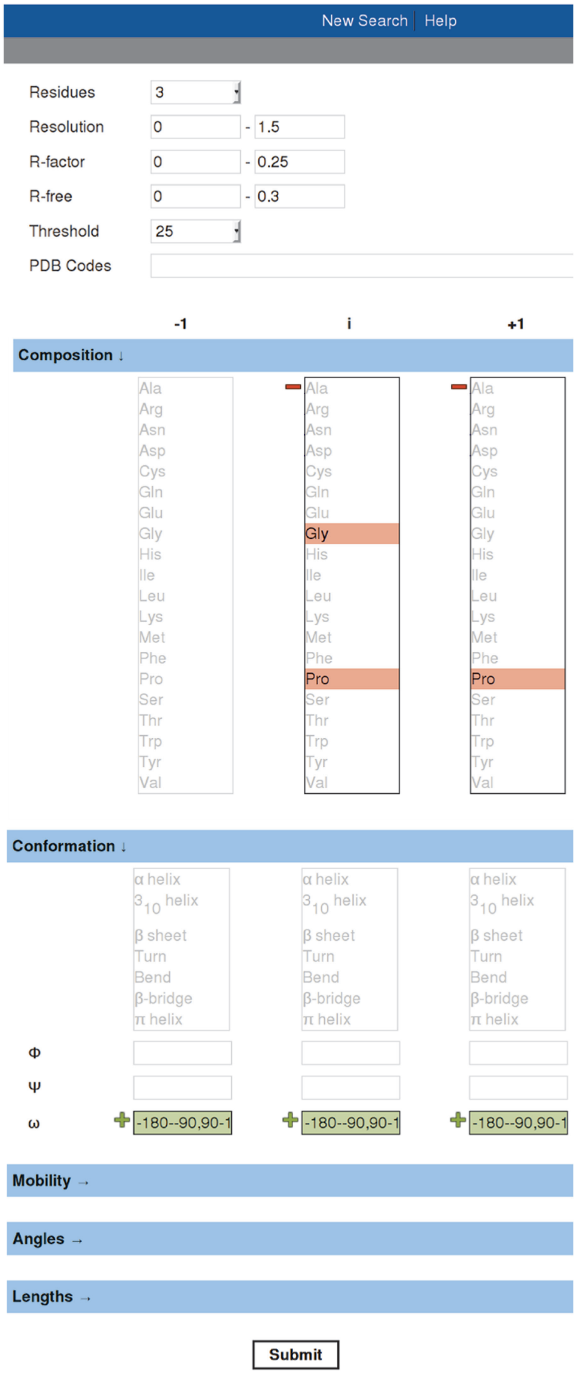


Figure 2. Extent and diversity of the database. The residue population of the PGD is shown as a function of resolution (**A**), amino-acid composition (**B**), and secondary-structure type (**C**). The population as a function of resolution is cumulative. At 1.0 Å resolution or better, the current PGD version contains 30 256 residues. Secondary-structure types are defined as follows: 'H' — α-helix; 'G' — 3_{10} helix; 'E' — β-strand; 'T' — hydrogen-bonded turn; 'S' — non-hydrogen-bonded turn; 'I' — π-helix; and 'B' — β-bridge. The 'I' type (π-helix) bar is too small to be visible, with only 687 observations.



Residues 3
Resolution 0 - 1.5
R-factor 0 - 0.25
R-free 0 - 0.3
Threshold 25
PDB Codes

Figure 3. Excerpt from a representative query. The query form defines a search for three-residue motifs that do not include Gly, Pro, or prePro residues at position i at 1.5 Å resolution or better. For residue composition, red highlights indicate excluded residues. For flexible-syntax boxes, green highlights indicate valid input, and an example of the flexible query syntax is visible: '–180–90,90–180' for ω, describing a search for *trans* peptides.

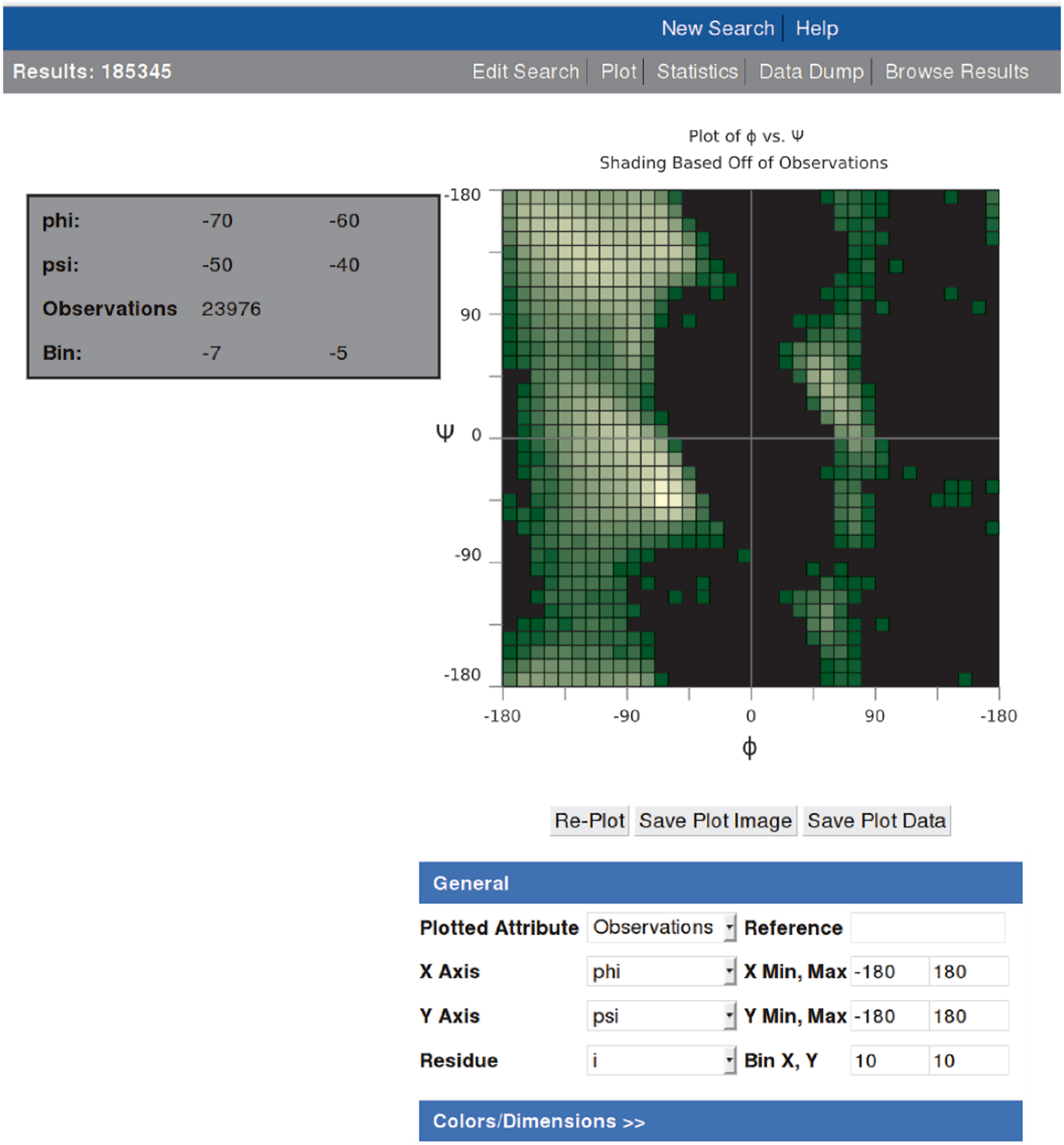
of their queries. A help window pops up for each section when entering data. Each of the criteria can be defined positively (e.g. Gly and Pro) or negatively (e.g. all but Gly, Pro).

At the top of the page is the length of the motif to be searched (from 1 to 10 residues), followed by protein-chain properties and residue properties. The protein-chain properties are the length of the motif, the resolution range for selecting crystal structures, the sequence-identity threshold, and specific PDB codes to search (defaults to the full PGD). Changing the motif length will cause the corresponding set of residue properties to appear.

The residue properties are grouped into five sections: composition, conformation, mobility, angles, and lengths. The composition section allows users to indicate any grouping of specific amino-acid types to search (i.e. with no limitation to predefined categories such as hydrophobic or acidic). The conformation section allows users to restrict searches to specific classes of DSSP-defined secondary structure, defined as follows: 'H' — α-helix; 'G' — 3_{10} helix; 'E' — β-strand; 'T' — hydrogen-bonded turn; 'S' — non-hydrogen-bonded turn; 'I' — π-helix; and 'B' — β-bridge. The long names are used on the search page, and the short names are used when space is limited (e.g. on the statistics page). The conformation section additionally offers options for more fine-grained conformational searches using ranges of φ, ψ, and the peptide planarity, ω. The mobility, angles, and lengths sections are collapsed by default to simplify the search page for first-time users and for those who only want to perform conformational searches; clicking the titles will expand them (other sections can be expanded or hidden in the same manner). The mobility section allows searching ranges of the three B-factors of the mainchain, sidechain and C_γ-atom (B^m, B^{sc} and B^γ, respectively). The angles are defined by three atoms and proceed in order from N- to C-terminus of the residue, with '–1' indicating an atom from the previous residue and '+1' indicating an atom from the next residue. The lengths are defined by two atoms and are otherwise named and searched identically to angles.

To allow for additional flexibility and convenience in searches, we made two enhancements beyond what is typically allowed in similar databases. First, we created a query syntax for ranges that allows multiple ranges to be specified (using commas), which enables searches wrapping around circular angles in either direction (search ranges must always be specified as negative to positive from left to right). This is quite useful for searches of conformations (the β region extends beyond $\psi = +180°/-180°$) or peptide planarity (which peaks at $+180°/-180°$). To make it difficult for users to create an invalid search, we also provide on-the-fly validation that highlights valid syntax in green and invalid syntax in red. Second, we created a special exclusion feature for selections (a green plus sign indicates when selections are included, and clicking it reverses the search to exclusion and displays a red minus sign) that allows users to easily exclude a small number of selections instead of tediously selecting almost all of them. This is useful for common cases like excluding Gly or Pro from a search.

Once a search is fully defined, clicking the 'Submit' button passes the query to the PGD, which immediately indicates that a search is in progress and displays results on the initial output page when the search is complete.



Results: 185345

phi:	-70	-60
psi:	-50	-40
Observations	23976	
Bin:	-7	-5

Plotted Attribute	Observations	Reference		
X Axis	phi	X Min, Max	-180	180
Y Axis	psi	Y Min, Max	-180	180
Residue	i	Bin X, Y	10	10

Figure 4. Excerpt from a representative output. The Ramachandran plot shows results of a search for three-residue motifs that do not include Gly, Pro or prePro residues at position i at 1.5 Å resolution or better with other settings left at their defaults. Coloration of the plot in green indicates the observation density in each bin, from low (dark) to high (light). The gray popup box on the left gives information for the pixel over which the cursor is placed. It is one of the most highly populated bins in the α region. The total result count is visible at the left edge of the top navigation bar.

The initial output page

Immediately following a search, the total number of results is reported in the upper left-hand corner and the numbers of results are displayed as a function of φ and ψ on an interactive Ramachandran plot (Figure 4). The plot is colored by observation density within 10° × 10° bins. To maximize the visual contrast, coloring uses a logarithmic scale derived from the plotted values. Moving the mouse cursor over any bin produces a JavaScript popup indicating the φ and ψ ranges and the observation count. The Ramachandran plot is not limited to displaying the number of observations but instead can show any of the PGD residue attributes, using colors (like a contour plot) or even on the X and Y axes (replacing φ/ψ). Attributes from any position of the search (i–4 to i + 5) can be plotted by changing the 'residue' parameter. Additionally, plots can be zoomed by changing the minima and maxima, and bin sizes can be modified. Further flexibility is available in the colors/dimensions section, hidden by default. To re-plot after changing any parameter, click the 'Re-Plot' button. Plots or the summary data used to create them (bin definitions, observation counts, averages and standard deviations for each attribute) are downloadable using the 'Save Plot Image' or 'Save Plot Data' butttons, respectively.

Additional tools and analysis

A feature of the PGD is that it enables analysis beyond its built-in capabilities by allowing users to download a complete set of search results. Clicking 'Data Dump' will prompt the download of a plain-text dump of the raw results for each matching motif in tab-separated value format, ideal for importing into other applications.

In addition to the summary data provided by the Ramachandran plot, the individual motifs found by the search are viewable online by clicking 'Browse Results' at the top of the page. Highlighting of each column and row under the mouse cursor eases comparison within a residue or attribute. To reduce load time and maximize responsiveness, pagination splits up the potentially large result sets.

The 'Statistics' link at the top of the page leads to a page of summary statistics about residue i, including a breakdown of observations by amino-acid type, secondary-structure types, and the average backbone covalent geometry. Scrolling a mouse cursor over the covalent-geometry values produces a pop-up window that displays the standard deviations and ranges. Automatic highlighting of the column and row under the mouse cursor eases comparisons within residue types and attributes.

EXAMPLES

The PGD enables searches of conformational and geometric space in a powerful, flexible manner that allows for a wide variety of uses, from understanding large-scale patterns in protein structure to analyzing the significance and/or rarity of a feature in an individual structure. Here, we describe three examples from papers published by our group using the PGD that illustrate its two primary aspects of conformational and geometric searching as well as the connection between them.

Conformational searching

The first example (18) is a large-scale analysis of protein conformation that asked a simple question: What linear groups with repeating φ,ψ pairs exist in proteins. To answer this question, we used the PGD to search for well-defined ($B^m < 25\,Å^2$) three-residue segments from structures solved at 1.2 Å resolution or better. At this resolution, the atomic positions and thus the torsion angles have high accuracy so if linear groups are truly tightly grouped, they should be observed as such. To ensure a maximally representative result, we chose the 25% sequence-identity threshold and included all amino-acid types but only *trans* peptides (all three ω values limited to '−180–90,90–180'). To identify linear groups in specific regions, we required all three residues to be in the same 20° × 20° box and systematically searched all such boxes using a 10° sliding window. We found that only three true clusters of linear groups exist in proteins: the right-handed α-/3_{10}-helix, the β-strand (with no substantive difference between parallel and antiparallel), and the P_{II} helix (occupied by many nonproline residues, despite the misconception that only polyproline populates it). The 2.2_7 ribbon, π-helix, and left-handed α-/3_{10}-helical conformations only occur for isolated residues and rare short segments.

Geometric searching

The second example [see figure 4 of (19)] is a small-scale analysis investigating the commonality of a specific five-residue geometric motif. In glutathione reductase, a key active-site loop bridges two cysteines forming a redox-active disulfide bond. This loop has five consecutive residues with nonplanar peptide bonds, and intriguingly, they are all bent in the same direction with a summed deviation of 55° across the pentapeptide. We suspected this highly strained loop was involved in the enzyme's function. To find out how common such an ω-deviation was in proteins, we searched the PGD for all *trans* pentapeptides for which each residue was at least 5° away from planarity (ω delimiter: '–175–90,90–175'), using cutoffs of 1.2 Å resolution and 90% sequence identity. By downloading a data dump and creating a histogram of the net deviations, we discovered that the strained active-site loop of glutathione reductase was not just unique within its own structure but also nearly unique among all proteins, with only two other examples in the PGD.

Probing conformation/covalent geometry relationships

The third example (20) is a broad analysis of the variations in covalent geometry as a function of the backbone conformation. To perform this analysis, we searched the PGD for three-residue segments of *trans* peptides (ω values limited to '−180–90,90–180') with well-defined ($B^m < 25\,Å^2$) backbones from structures solved at 1.0 Å resolution or better, using a 90% sequence-identity threshold. We split the searches into five classes to define the differing behaviors of each class: Ile/Val, Gly, Pro, residues preceding Pro, and the remaining 16 residues. By downloading a data dump, we imported the data into Matlab for further in-depth analysis of the geometry trends. The results of this analysis were captured as a conformation-dependent geometry library that was used to show how accounting for these systematic relationships could improve the accuracy of homology modeling and crystallographic refinement.

CONCLUSIONS

As the examples illustrate, the ability to explore peptide geometry and conformation and their interrelationship can provide important insights into protein structure and function. The PGD is the only database to connect peptide geometry and conformation. Its highly flexible yet intuitive search interface will allow users to characterize principles of protein structure and to answer questions about details of protein structure that are often missed or ignored.

ACKNOWLEDGEMENTS

The authors would like to thank their beta testers, who made suggestions for improving the PGD. They would also like to thank all members of the development team at the Open Source Lab at Oregon State University who assisted with coding. They also thank Nan Wang, Peter Gross and Mike Marr for work on early versions of the PGD.

FUNDING

National Institutes of Health (R01-GM083136 to P.A.K.); National Science Foundation (MCB-9982727 to P.A.K.). Funding for open access charge: National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES

1. Lawson,C.L. (1996) An atomic view of the L-tryptophan binding site of trp repressor. *Nat. Struct. Biol.*, **3**, 986–987.
2. Dobson,R.C.J., Griffin,M.D.W., Devenish,S.R.A., Pearce,F.G., Hutton,C.A., Gerrard,J.A., Jameson,G.B. and Perugini,M.A. (2008) Conserved main-chain peptide distortions: a proposed role for Ile203 in catalysis by dihydrodipicolinate synthase. *Protein Sci.*, **17**, 2080–2090.
3. Merritt,E.A., Kuhn,P., Sarfaty,S., Erbe,J.L., Holmes,R.K. and Hol,W.G. (1998) The 1.25 A resolution refinement of the cholera toxin B-pentamer: evidence of peptide backbone strain at the receptor-binding site. *J. Mol. Biol.*, **282**, 1043–1059.
4. Davis,I.W., Leaver-Fay,A., Chen,V.B., Block,J.N., Kapral,G.J., Wang,X., Murray,L.W., Arendall,W.B., Snoeyink,J., Richardson,J.S. *et al.* (2007) MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.*, **35**, W375–W383.
5. Esposito,L., Vitagliano,L., Zagari,A. and Mazzarella,L. (2000) Experimental evidence for the correlation of bond distances in peptide groups detected in ultrahigh-resolution protein structures. *Protein Eng.*, **13**, 825–828.
6. Laidig,K.E. and Cameron,L.M. (1993) What happens to formamide during C—N bond rotation? Atomic and molecular energetics and molecular reactivity as a function of internal rotation. *Can. J. Chem.*, **71**, 872–879.
7. Karplus,P.A. (1996) Experimentally observed conformation-dependent geometry and hidden strain in proteins. *Protein Sci.*, **5**, 1406–1420.
8. Kleywegt,G.J. (1999) Recognition of spatial motifs in protein structures. *J. Mol. Biol.*, **285**, 1887–1897.
9. Ananthalakshmi,P., Kumar,C.K., Jeyasimhan,M., Sumathi,K. and Sekar,K. (2005) Fragment Finder: a web-based software to identify similar three-dimensional structural motif. *Nucleic Acids Res.*, **33**, W85–W88.
10. Samson,A.O. and Levitt,M. (2009) Protein segment finder: an online search engine for segment motifs in the PDB. *Nucleic Acids Res.*, **37**, D224–D228.
11. Sheik,S.S., Ananthalakshmi,P., Bhargavi,G.R. and Sekar,K. (2003) CADB: Conformation Angles DataBase of proteins. *Nucleic Acids Res.*, **31**, 448–451.
12. Golovin,A. and Henrick,K. (2008) MSDmotif: exploring protein sites and motifs. *BMC Bioinformatics*, **9**, 312.
13. Wang,G. and Dunbrack,R.L. (2003) PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.
14. Laskowski,R.A., MacArthur,M.W., Moss,D.S. and Thornton,J.M. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.*, **26**, 283–291.
15. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
16. Lovell,S.C., Davis,I.W., Arendall,W.B. III, de Bakker,P.I.W., Word,J.M., Prisant,M.G., Richardson,J.S. and Richardson,D.C. (2003) Structure validation by Ca geometry: Π, ψ and Cβ deviation. *Proteins: Struct. Func. Genet.*, **50**, 437–450.
17. Cock,P.J.A., Antao,T., Chang,J.T., Chapman,B.A., Cox,C.J., Dalke,A., Friedberg,I., Hamelryck,T., Kauff,F., Wilczynski,B. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
18. Hollingsworth,S.A., Berkholz,D.S. and Karplus,P.A. (2009) On the occurrence of linear groups in proteins. *Protein Sci.*, **18**, 1321–1325.
19. Berkholz,D.S., Faber,H.R., Savvides,S.N. and Karplus,P.A. (2008) Catalytic cycle of human glutathione reductase near 1 A resolution. *J. Mol. Biol.*, **382**, 371–384.
20. Berkholz,D.S., Shapovalov,M.V., Dunbrack,R.L. and Karplus,P.A. (2009) Conformation dependence of backbone geometry in proteins. *Structure*, **17**, 1316–1325.
21. Antonyuk,S.V., Strange,R.W., Sawers,G., Eady,R.R. and Hasnain,S.S. (2005) Atomic resolution structures of resting-state, substrate- and product-complexed Cu-nitrite reductase provide insight into catalytic mechanism. *Proc. Natl Acad. Sci. USA*, **102**, 12041–12046.