BIOINFORMATICS APPLICATION NOTE

Vol. 31 no. 2 2015, pages 292–294
doi:10.1093/bioinformatics/btu617

Databases and ontologies

Advance Access publication September 23, 2014

PHI-DAC: protein homology database through dihedral angle conservation

Noa Maatuk, Yitav Glantz-Gashai, Maya Rotman, Meirav Baydany, Gennadiy Fonar, Amir Shechvitz, Karin Shemer, Aviva Peleg, Eli Reuveni and Abraham O. Samson*

Faculty of Medicine in the Galilee, Bar Ilan University, 13100 Safed, Israel

Associate Editor: Janet Kelso

ABSTRACT

Finding related conformations in the Protein Data Bank is essential in many areas of bioscience. To assist this task, we designed a dihedral angle database for searching protein segment homologs. The search engine relies on encoding of the protein coordinates into text characters representing amino acid sequence, φ and ψ dihedral angles. The search engine is advantageous owing to its high speed and interactive nature and is expected to assist scientists in discovering conformation homologs and evolutionary kinship. The search engine is fast, with query times lasting a few seconds, and freely available at http://tarshish.md.biu.ac.il/~samsona

Contact: avraham.samson@biu.ac.il

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 28, 2013; revised on June 30, 2014; accepted on September 12, 2014

1 INTRODUCTION

Over the past years, structural data in the Protein Data Bank (PDB) has grown exponentially. At the present time, ~92 000 structures are available in the PDB (Berman *et al.*, 2002). The data abundance makes it difficult to navigate the information, particularly when one tries to retrieve segment motifs (Levitt, 1992). To find segment motifs in the PDB structures, several computational engines have been designed. One such engine is SPASM, which finds spatial motifs consisting of arbitrary main-chain and side-chain conformation in a database of protein structures (Kleywegt, 1999). An additional search engine named Fragment Finder was designed to identify similar structural motifs based on backbone dihedral angles (Ananthalakshmi *et al.*, 2005). Another engine, PAST, is based on translation- and rotation-invariant representation of protein backbone α-dihedral angles (Taubig *et al.*, 2006). Both PAST and Fragment Finder suffer from dihedral angle discretization of 5°, which coarse grains the results, and preclude amino acid sequence, which prevents a comprehensive search. Also noteworthy is the Dali Server, which enables the user to browse for homologs in Cartesian space (Holm *et al.*, 2008). None of these engines allows a combined search using amino acid sequence and dihedral angles in a rapid manner.

*To whom correspondence should be addressed.

Here we describe an online search engine that rapidly finds protein segments based on amino acid sequence and φ and ψ dihedral angles. The search engine, named Protein Homology through Dihedral Angle Constraints (PHI-DAC), is advantageous owing to its speed, generality and simplicity. It is expected to be helpful to the scientific community by easing the identification of conformation homologs and distilling useful information from the PDB.

2 METHODS

Protein structure representation. To describe the 3D conformation of a protein we use φ and ψ dihedral angles. These angles have the advantage of being invariant to translation and rotation of the protein structure in a coordinate system. The angles are calculated using DSSP (Kabsch and Sander, 1983) and encoded into a structural alphabet (Levitt, 1992) represented by text characters from 0 to 359°. This transforms the information of the 3D backbone conformation from all protein structures contained in the PDB into sequences of text that are easily searchable.

Construction of the PHI-DAC database. All PDB files containing protein structure are handled as separate entries and included in PHI-DAC. As of January 2014, a total of ~96 000 files describe protein structures. Computation of the PHI-DAC database, given the φ and ψ dihedral angles and amino acid sequences, takes ~1 min on a standard PC (2.7 GHz). The size of the database is 174 MB, and it can be held in working memory, making all calculations extremely fast. The search engine is hosted on a 2.7 GHz PC with 2 GB random access memory operating under Linux. Update of the PHI-DAC database is expected monthly.

3 RESULTS

PHI-DAC database. In Table 1, an example entry of the PHI-DAC database is shown. The entry is composed of four pipe delimited fields. The first and second fields contain the PDB identification (ID) and the amino acid sequence, respectively. Chain IDs are referenced to the DSSP files, and cysteines are in lowercase letters. Duplicate lowercase letters form disulfide bridges, 'a' with 'a', 'b' with 'b', etc. Exclamation points denote a chain break or a new polypeptide chain. The third and fourth fields contain the φ and ψ dihedral angles encoded in double characters. The first character represents units of 10 (A = 0, B = 10 . . . S = 180, and a = –0, b = –10 . . .) and the second character represents units of 1. Thus, 'A0' is 0°, 'A1' is +1°, 'A2' is +2°, 'S0' is +180° (or –180°), 'a1' is –1° (or 359°),

Table 1. Example entry of PHI-DAC database

```
1l4w|IVaHTTATSPISAVTbPPGENLaYRKMWcDAFcSSRGKVV
ELGbAATdPSKKPYEETdeSTDKeNPHPKQRPG!EERGWKHWVYY
TCCPDTPYLDITEE|A0i3b6i2a8g1c1k4l3i2l7i5d1e0i5
j0l8k9l1k4f9l0i1i4g2k2c4e5h0e0J6l6k4i8d8k7b9j
9j9e0f2i4c4f8f1f8f9o2l0j0j0m0o9R6c6f6b0i2j1e7
f7e9o3D0j4g8j9j6j3l1h6R7j1k9A0A0i4G1i0f8d9e0e
0c3i2i2l1O5d5i9j6f9k8h3i2e1f4b7b0l0|Q9Q4P0R5Q
2J8n4d5l5b5N6P2L6Q6O1K6n3b4p3J4A2N3L1P0I3N9N0
c3M0I3p1p0o6k6m8j0a4O6P0M3M1N4R8L9R9p2F3K0a7N
7B3o1n0C9b7Q6L5K8L4P3E9P4M3F6B1G0R0P5l6p1K9m8
F1A0A0o1k2L8g6g2K8M4R0J5L8M6M7n6M9p8s0F1O4K0O
4L3m4O3I5A0
```

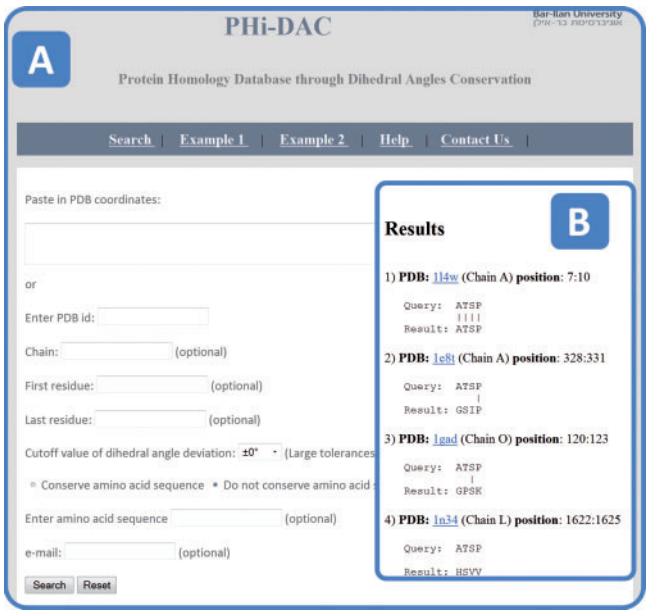

Fig. 1. The PHI-DAC search engine. The search engine finds protein segment homologs based on φ and ψ dihedral angles and amino acid sequence. (**A**) Search parameters include PDB coordinates, PDB and chain ID, as well as residue range. (**B**) The result page of PHI-DAC lists the segments that matched the query. Shown here is an example query for segment 7–10 of chain A of PDB ID 1L4W with a torsion angle tolerance of 8°. The results are listed in order of the RMSD of dihedral angles between the query and result. Notice sequence identity indicated by pipes

and so on. This protein backbone representation is concise and accurate. The database is available for download.

Querying PHI-DAC. To perform a search for similar protein backbone conformations in the PDB, the web interface offers the following search parameters: PDB ID, chain, first and last residue and tolerance. The tolerance is a range of neighboring dihedral angle intervals in 1° increments. The dihedral angle discretization of 1° is superior to Fragment Finder and PAST and allows for more accurate results. The tolerance is limited to ±30°, so as to provide meaningful spatial superposition. A screenshot of the query interface is shown in Figure 1. Supplementary Table S1 summarizes the run times and the work range of the server in terms of peptide length, complexity of fold and dihedral angle deviation. Note that the bottleneck of the server is the dihedral angle deviation.

Interpreting the results. Following a query submission, the result table should be loaded automatically within few seconds. A screenshot of the result table of the example query is shown in Figure 1. The result page contains the PDB ID, chain, position and the amino acid sequence. The PDB IDs are hyperlinked to the respective PDB file. The results are classified according to the dihedral angle root-mean-square deviation (RMSD) between the query segment and the result segment (low to high).

Example queries. Two example queries are readily accessible on the Web site. In the first example query the coordinates of a small bent helix in PDB ID 9XIM (residues 121–127) are used. Searching with the φ and ψ dihedral angles by using a tolerance of ±3° and without conserving the amino acid sequence leads to 14 matching segments belonging to 14 different PDB entries. In the second example query, the coordinates of the β-hairpin segment 7–10 of chain A of PDB ID 1L4W is used. Querying with the φ and ψ dihedral angles by using a tolerance of ±8° and without conserving the amino acid sequence leads to 13 matching segments belonging to 13 different PDB entries.

Local versus global homology. PHI-DAC is based on local dihedral angle homology and is not designed to detect global spatial similarities. As such, PHI-DAC does not detect similarity between large SCOP or CATH folds with perfect superimposition in the core and major differences in the loops. For spatial superimposition, SCOP, CATH and Dali should be used.

4 CONCLUSION

Our method of encoding the backbone φ and ψ dihedral angles into text characters has proven to be a fast solution for answering queries about local structural similarities in the PDB. Compared with SPASM, PHI-DAC shows similar results, while being much faster and addressing a different question compared with Dali. Therefore, compared with Dali, we consider PHI-DAC to be a valuable tool for the fast detection of protein segments based on dihedral angle conservation.

ACKNOWLEDGEMENT

We thank Prof. Haim Breitbart for helpful comments.

Funding: This research was supported by the Katz Foundation and a Marie-Curie CIG grant 322113 (to A.O.S.).

Conflict of interest: none declared.

REFERENCES

Ananthalakshmi,P. *et al.* (2005) Fragment finder: a web-based software to identify similar three-dimensional structural motif. *Nucleic Acids Res.*, **33**, W85–W88.

Berman,H.M. *et al.* (2002) The protein data bank. *Acta Crystallogr. D Biol. Crystallogr.*, **58**, 899–907.

Holm,L. *et al.* (2008) Searching protein structure databases with DaliLite v.3. *Bioinformatics*, **24**, 2780–2781.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
Kleywegt,G.J. (1999) Recognition of spatial motifs in protein structures. *J. Mol. Biol.*, **285**, 1887–1897.
Levitt,M. (1992) Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol.*, **226**, 507–533.
Taubig,H. *et al.* (2006) PAST: fast structure-based searching in the PDB. *Nucleic Acids Res.*, **34**, W20–W23.