doi:10.1093/nar/gkq522

ProteinDBS v2.0: a web server for global and local protein structure search

Chi-Ren Shyu[1,2,*], Bin Pang[1], Pin-Hao Chi[2], Nan Zhao[1], Dmitry Korkin[1,2] and Dong Xu[1,2]

[1]Informatics Institute and [2]Department of Computer Science, University of Missouri, Columbia, MO 65211, USA

Received February 24, 2010; Revised April 26, 2010; Accepted May 24, 2010

ABSTRACT

ProteinDBS v2.0 is a web server designed for efficient and accurate comparisons and searches of structurally similar proteins from a large-scale database. It provides two comparison methods, global-to-global and local-to-local, to facilitate the searches of protein structures or substructures. ProteinDBS v2.0 applies advanced feature extraction algorithms and scalable indexing techniques to achieve a high-running speed while preserving reasonably high precision of structural comparison. The experimental results show that our system is able to return results of global comparisons in seconds from a complete Protein Data Bank (PDB) database of 152 959 protein chains and that it takes much less time to complete local comparisons from a non-redundant database of 3276 proteins than other accurate comparison methods. ProteinDBS v2.0 supports query by PDB protein ID and by new structures uploaded by users. To our knowledge, this is the only search engine that can simultaneously support global and local comparisons. ProteinDBS v2.0 is a useful tool to investigate functional or evolutional relationships among proteins. Moreover, the common substructures identified by local comparison can be potentially used to assist the human curation process in discovering new domains or folds from the ever-growing protein structure databases. The system is hosted at http://ProteinDBS.rnet.missouri.edu.

INTRODUCTION

The great demand for an efficient and accurate search engine for 3D protein structures has continued to rise due to the dramatic increase in protein structural data and the role of protein structures in biological findings (1). The number of known protein structures in the primary structural database, the Protein Data Bank (PDB), had reached 63 271 (~152 959 protein chain structures) as of 16 February 2010, and is expected to continue growing at a high rate. The most important and difficult task in handling such a large number of protein structures is to develop an efficient and accurate tool for fast comparison between a new structure and all existing ones in the database, so as to discover potential biological connections. To assist in this task, a high-throughput and accurate structural comparison method is essential. Traditional comparison methods, such as DALI (2) and CE (3), are based on the calculation of a distance matrix of residues, which can provide accurate alignment but are usually computationally expensive.

In recent years, approaches have been developed to improve the performance of structural comparison and search. Fast web servers, including TOPSCAN (4), YAKUSA (5), 3D-Blast (6) and iSARST (7) map protein structures into 1D sequences and then use various sequence alignment methods to align two structures. These approaches exhibit good efficiency; however, 1D representations of 3D structures potentially lose details of structural topologies, which could lead to lower accuracy than the accurate structural comparison methods (6).

To meet the challenges of strict efficiency and accuracy requirements from the large-scale protein structure database, ProteinDBS was initially developed in 2003 to provide the community a real-time web server for searching globally similar protein structures (8). The first generation of ProteinDBS has been widely used by the community and recognized by the 3 September 2004, issue of Science (9). During these years, ProteinDBS has been continuously improved in performance and service to keep it as a useful resource to study protein structures.

In the new version of ProteinDBS, major advancements include local structural comparison as well as biologist-friendly query interfaces and visualization tools. In contrast to the global comparison, which tries to superimpose most of the corresponding backbone atoms from two proteins, the local comparison seeks to find all common substructures between two proteins. As an example, gap-free common substructures are usually

*To whom correspondence should be addressed. Tel: +1 573 882 3884; Fax: +1 573 884 8709; Email: shyuc@missouri.edu

linked by coils of different lengths in a protein structure family. These multiple common substructures, different from canonical secondary structures, might be used not only for the discovery of new domains or folds in the structure database but also for the identification of functional and evolutionary relationships of protein structures since they are more conserved than other regions (10–12). However, aligning protein substructures is known to be a non-deterministic polynomial (NP) time-hard problem, and the existing methods are rarely designed to handle such kind of problem. The problem becomes more challenging as one considers the growing rates at which new protein structures are being added to the database. Hence, the main goal of ProteinDBS v2.0 is to tackle these issues and equip the community with user-friendly tools that can deliver efficient and accurate results for protein structure comparison and search.

ProteinDBS v2.0 has been optimized in the following aspects:

(i) The global comparison method introduces a novel knowledge-based feature extraction as well as online database indexing (13).
(ii) The newly designed approach for local comparisons applies information retrieval algorithms to extract frequently occurring substructure patterns and utilizes an index tree to efficiently manage the data (14).
(iii) The database of global comparisons is being updated weekly from the PDB website, and the database of local comparisons has been generated using the latest release of SCOP 1.75 (15) and the non-redundant protein data set PDBSelect (16).
(iv) The efficiency of ProteinDBS has been significantly enhanced to meet the requirements of large-scale protein database searching.

To our knowledge, ProteinDBS v2.0 is the only server that can simultaneously support large-scale comparison and search of globally and locally similar protein structures.

MATERIALS AND METHODS

The system architecture of ProteinDBS v2.0, as shown in Figure 1, contains five modules: (i) protein structure database management; (ii) data pre-processing; (iii) query interface; (iv) distributed search engine; and (v) retrieval results visualization. A system tutorial can be viewed at the ProteinDBS web site.

Protein structure database management

ProteinDBS v2.0 maintains two independent databases of protein structures for global and local comparisons. The database of global comparisons is updated weekly and newly added protein structures are automatically downloaded from the PDB ftp site (ftp://ftp.wwpdb.org/pub/pdb/). For each new protein structure, a 2D distance matrix is generated from 3D coordinates of the protein chains. The distance matrices are empirically proven capable of representing global protein structural topologies. From the distance matrices, 33 features are then extracted, and a tree structure, an M-Tree (17), is utilized to index the multi-dimensional data.

The database of local comparisons is a non-redundant data set of protein chains selected from PDBSelect (16) and SCOP v1.75 (15). When new data set is released, substructure units, defined as continuous fragments of backbone with fixed length, are first identified from each protein in the data set using a sliding window. Our assumption is that a protein containing similar substructure units should be further investigated to find long common substructures. In order to efficiently search proteins with similar substructures, the system first organizes structurally similar substructure units into a cluster and selects a representative for each cluster. The representative is assigned a label called ‘term’ in our server. The system then maps the protein structure into a series of terms by comparing the substructure units with the pre-defined substructure representative of each cluster. Finally, the system utilizes an M-Tree to index the terms of the entire database of proteins to facilitate fast searches using information retrieval techniques.

Query interface

There are two types of query methods, as shown in the top block of Figure 1: local-to-local and global-to-global search. Both methods feature ‘query by ID’ and ‘query by structure example’. Using an internet browser, a user can upload a new protein chain structure in PDB format or provide a PDB ID contained in the protein database to find similar protein structures.

Data pre-processing and search engine

The global-to-global search, as mentioned previously, first maps the query protein structures into a 2D distance matrix and extracts features from the distance matrix. In this way, the query protein can be represented by a data point in the feature space populated by the entire protein database. Thus, one-against-all global comparison is analogous to searching nearest neighbors in feature space, and such a search can be completed in real time.

For the local-to-local search, the system first extracts substructure units of the query protein and then clusters them into groups. From the index of database terms, the system finds candidate proteins for comparison and filters out those proteins without common substructures. To achieve accurate substructure comparison, the system deploys a coarse-to-fine strategy to align the query protein and a database protein. Specifically, the system first finds relevant matches at the substructure level with a customized dynamic programming algorithm (14) and then refines the substructure alignment at the atomic level. This two-level alignment framework is a tradeoff allowing the system to achieve high efficiency without sacrificing accuracy of results.

Due to page limitations, interested readers are referred to (13,14) for further discussion of the detailed algorithms.

Retrieval results visualization

The global comparison retrieval results for a query chain 1o7j_A are shown in Figure 2. A set of top-ranked

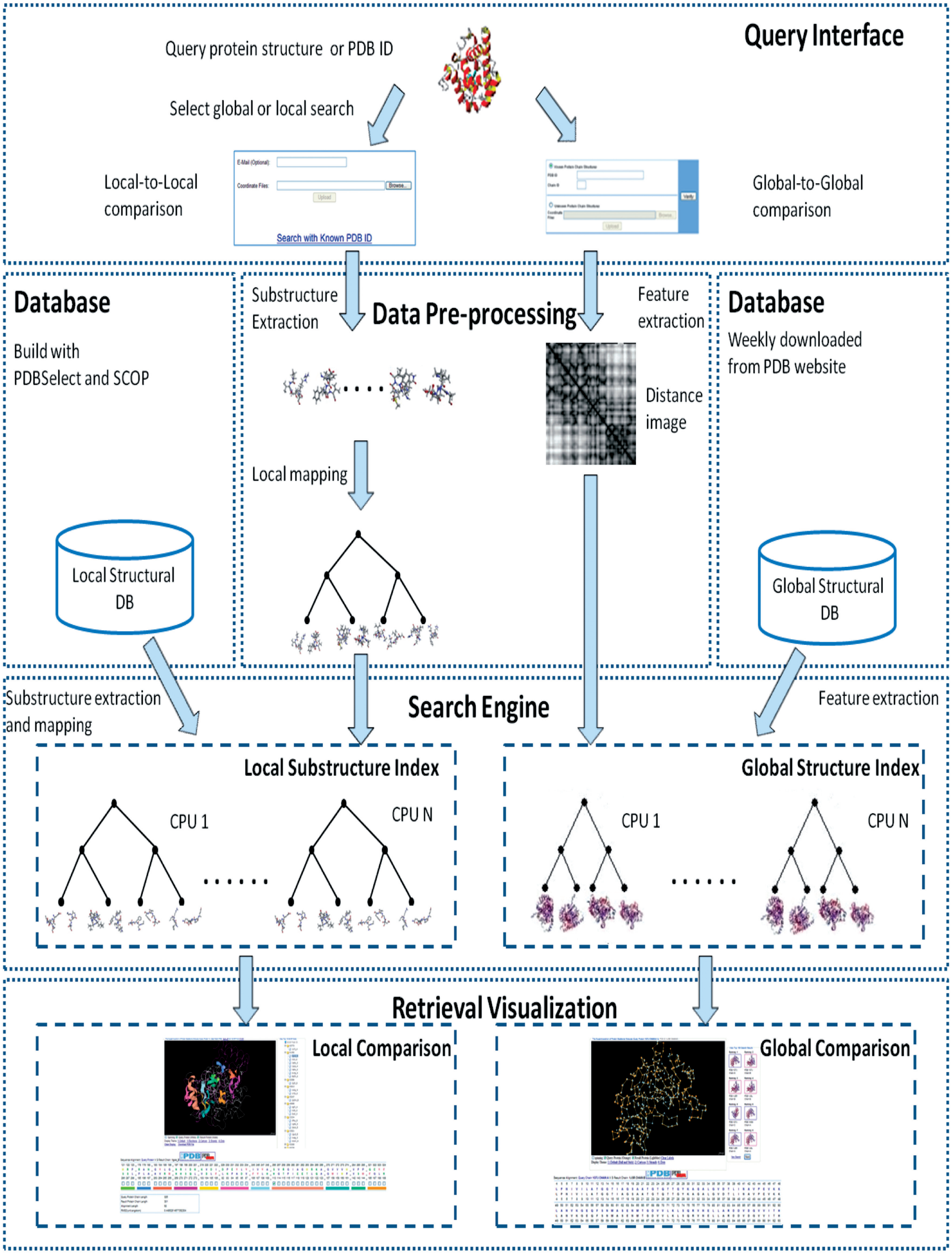
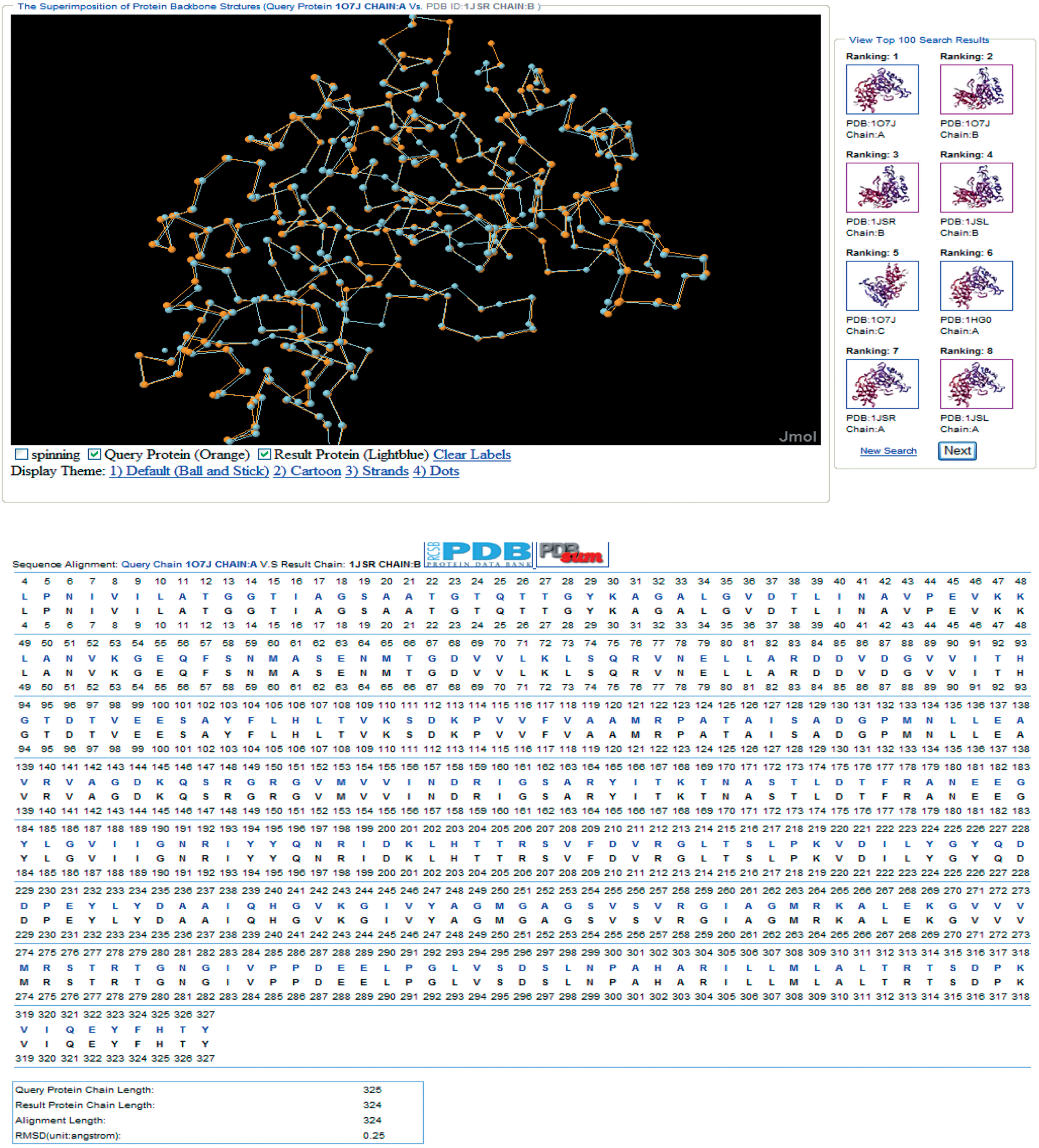


Figure 1. ProteinDBS v2.0 system architecture which has five modules (i) query interface; (ii) protein structure data management; (iii) data pre-processing; (iv) search engines; and (v) retrieval result visualization.

structures is returned to the user, eight at a time. To visualize the quality of the search results, a 3D superimposition of the query and the top-retrieval result are displayed to the user. The user can select any of the ranked results from the top-right panel. Figure 2 presents a new interface for the superimposition view of the query protein chain and the top-ranked result, 1jsr_B, which is generated by clicking on the thumbnail image on the top-right panel. The sequence alignment result is also displayed to the user with root mean square deviation (RMSD) and alignment length values.



The Superimposition of Protein Backbone Strctures (Query Protein 1O7J CHAIN:A Vs. PDB ID:1JSR CHAIN:B)

□ spinning ☑ Query Protein (Orange) ☑ Result Protein (Lightblue) Clear Labels
Display Theme: 1) Default (Ball and Stick) 2) Cartoon 3) Strands 4) Dots

View Top 100 Search Results

Ranking: 1 PDB:1O7J Chain:A	Ranking: 2 PDB:1O7J Chain:B
Ranking: 3 PDB:1JSR Chain:B	Ranking: 4 PDB:1JSL Chain:B
Ranking: 5 PDB:1O7J Chain:C	Ranking: 6 PDB:1HG0 Chain:A
Ranking: 7 PDB:1JSR Chain:A	Ranking: 8 PDB:1JSL Chain:A

Sequence Alignment: Query Chain 1O7J CHAIN:A V.S Result Chain: 1JSR CHAIN:B

```
  4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48
  L   P   N   I   V   I   L   A   T   G   G   T   I   A   G   S   A   A   T   G   T   Q   T   T   G   Y   K   A   G   A   L   G   V   D   T   L   I   N   A   V   P   E   V   K   K
  L   P   N   I   V   I   L   A   T   G   G   T   I   A   G   S   A   A   T   G   T   Q   T   T   G   Y   K   A   G   A   L   G   V   D   T   L   I   N   A   V   P   E   V   K   K
  4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48

 49  50  51  52  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90  91  92  93
  L   A   N   V   K   G   E   Q   F   S   N   M   A   S   E   N   M   T   G   D   V   V   L   K   L   S   Q   R   V   N   E   L   L   A   R   D   D   V   D   G   V   V   I   T   H
  L   A   N   V   K   G   E   Q   F   S   N   M   A   S   E   N   M   T   G   D   V   V   L   K   L   S   Q   R   V   N   E   L   L   A   R   D   D   V   D   G   V   V   I   T   H
 49  50  51  52  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90  91  92  93

 94  95  96  97  98  99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138
  G   T   D   T   V   E   E   S   A   Y   F   L   H   L   T   V   K   S   D   K   P   V   V   F   V   A   A   M   R   P   A   T   A   I   S   A   D   G   P   M   N   L   L   E   A
  G   T   D   T   V   E   E   S   A   Y   F   L   H   L   T   V   K   S   D   K   P   V   V   F   V   A   A   M   R   P   A   T   A   I   S   A   D   G   P   M   N   L   L   E   A
 94  95  96  97  98  99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138

139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183
  V   R   V   A   G   D   K   Q   S   R   G   R   G   V   M   V   V   I   N   D   R   I   G   S   A   R   Y   I   T   K   T   N   A   S   T   L   D   T   F   R   A   N   E   E   G
  V   R   V   A   G   D   K   Q   S   R   G   R   G   V   M   V   V   I   N   D   R   I   G   S   A   R   Y   I   T   K   T   N   A   S   T   L   D   T   F   R   A   N   E   E   G
139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183

184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228
  Y   L   G   V   I   I   G   N   R   I   Y   Y   Q   N   R   I   D   K   L   H   T   T   R   S   V   F   D   V   R   G   L   T   S   L   P   K   V   D   I   L   Y   G   Y   Q   D
  Y   L   G   V   I   I   G   N   R   I   Y   Y   Q   N   R   I   D   K   L   H   T   T   R   S   V   F   D   V   R   G   L   T   S   L   P   K   V   D   I   L   Y   G   Y   Q   D
184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228

229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273
  D   P   E   Y   L   Y   D   A   A   I   Q   H   G   V   K   G   I   V   Y   A   G   M   G   A   G   S   V   S   V   R   G   I   A   G   M   R   K   A   L   E   K   G   V   V   V
  D   P   E   Y   L   Y   D   A   A   I   Q   H   G   V   K   G   I   V   Y   A   G   M   G   A   G   S   V   S   V   R   G   I   A   G   M   R   K   A   L   E   K   G   V   V   V
229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273

274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318
  M   R   S   T   R   T   G   N   G   I   V   P   P   D   E   E   L   P   G   L   V   S   D   S   L   N   P   A   H   A   R   I   L   L   M   L   A   L   T   R   T   S   D   P   K
  M   R   S   T   R   T   G   N   G   I   V   P   P   D   E   E   L   P   G   L   V   S   D   S   L   N   P   A   H   A   R   I   L   L   M   L   A   L   T   R   T   S   D   P   K
274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318

319 320 321 322 323 324 325 326 327
  V   I   Q   E   Y   F   H   T   Y
  V   I   Q   E   Y   F   H   T   Y
319 320 321 322 323 324 325 326 327
```

Query Protein Chain Length:	325
Result Protein Chain Length:	324
Alignment Length:	324
RMSD(unit:angstrom):	0.25

Figure 2. ProteinDBS retrieval results visualization for global-to-global comparison. The top-left panel shows a superimposed view of a query protein chain and a selected result chain from the ranked list in the top-right panel (the third-ranked protein chain in this figure). Users can check 'Result Protein' and/or 'Query Protein' to view the retrieved chain and/or the query chain. By clicking on a hyperlink below the top-left panel, users can view different display themes. The lower panel displays the structure alignment results represented by sequences with the RMSD and alignment length.

For a global structure search, users can anticipate real-time results. Local structure searches, however, usually take minutes, dependent on the size of the query protein. Our system provides two options for the users: (i) the system will return a session ID for the query along with an estimated execution time after the query protein structure has been uploaded. The user can then bookmark the link of the session ID and check back with the resulting page a few minutes later. (ii) If the user is willing to provide an email address when the query protein structure is uploaded, the system will send ranked results to the user's email account.

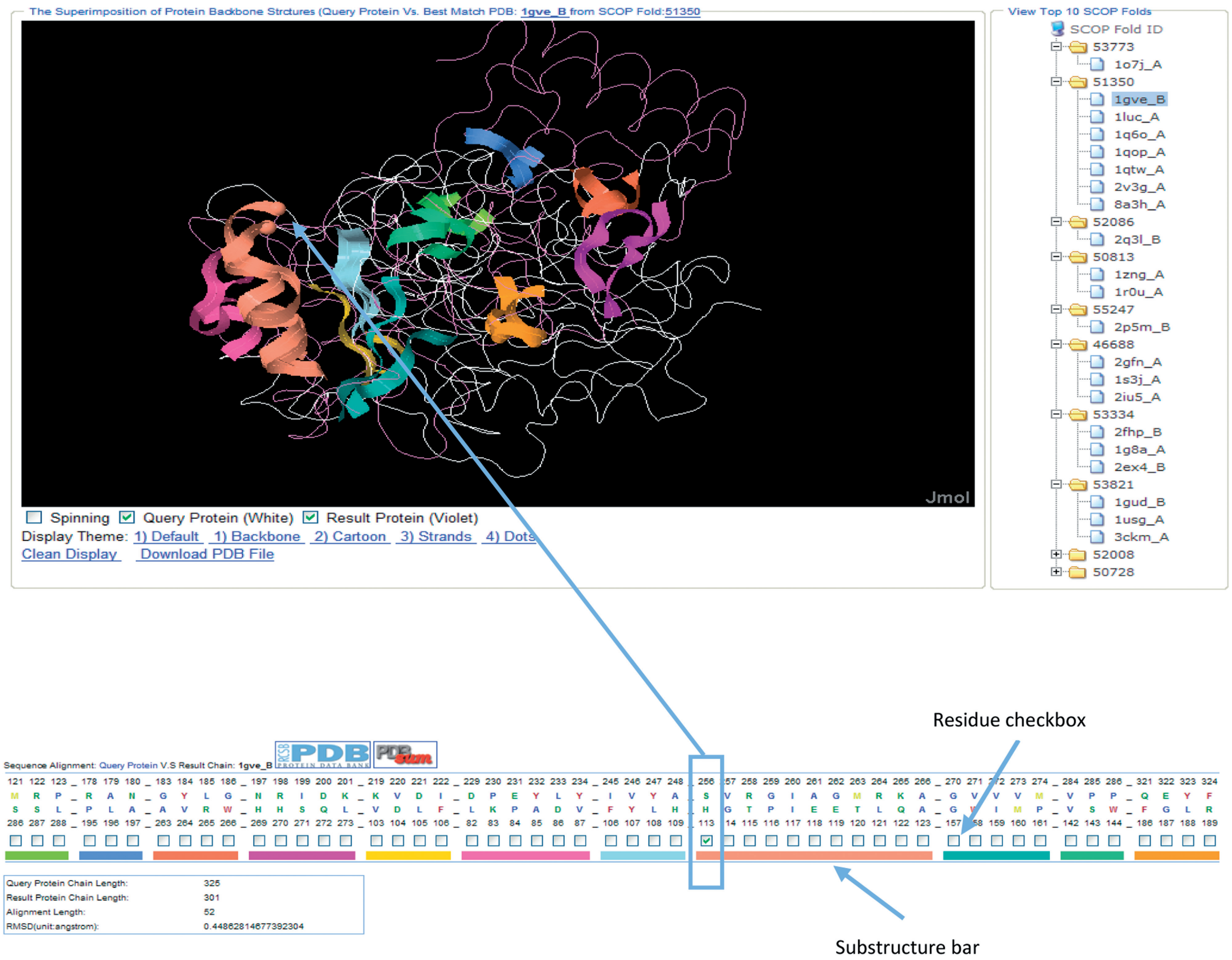
In the query page for local-to-local search, users can perform various search options by specifying (i) the view mode of the results; (ii) the session ID that was assigned after the query was submitted; and (iii) a threshold for substructure sizes. The local comparison method supports three types of result browsing modes: M_1, in which the top 10 SCOP folds with the best matched protein structure are shown; M_2, in which the top 100 matched protein structures from different SCOP folds are displayed; and M_3, in which the top 10 SCOP folds are presented with all matched proteins from the same fold.

Figure 3 shows an example of M_3, which organizes the retrieval results using a tree-view on the top-right panel. The top-left panel presents the superimposition of the query protein, 1o7j_A, and one of the top matched database proteins, 1gve_B, with a substructure size threshold of >3 residues. The common substructures are highlighted with different colors. The lower panel shows the sequence alignment result with RMSD and alignment length values. The users can use the 'residue checkbox' and the 'substructure bar' under the residues to interact with the 3D superimposition view. The superimposition is shown with all the qualified substructures at the beginning. Each substructure pair is differentiated with different colors in the 3D view and sequence alignment. When investigating a specific substructure, the users first use the hyperlink 'Clear Display' to hide all the substructures and then click on the 'substructure bar' to show the substructures in the 3D view. Similarly, clicking on the 'residue checkbox' will highlight one designated residue.

In addition, users can specify different display themes, such as backbone, cartoon, strand and dots, by clicking on the corresponding label. All aligned protein structures can be downloaded from the result pages.

Performance evaluation

Two major performance evaluations have been conducted for ProteinDBS, namely retrieval accuracy and efficiency. If the top-ranked results are from the same structural



The Superimposition of Protein Backbone Strctures (Query Protein Vs. Best Match PDB: 1gve_B from SCOP Fold:51350

□ Spinning ☑ Query Protein (White) ☑ Result Protein (Violet)

Display Theme: 1) Default 1) Backbone 2) Cartoon 3) Strands 4) Dots

Clean Display Download PDB File

View Top 10 SCOP Folds

SCOP Fold ID
- 53773
  - 1o7j_A
- 51350
  - 1gve_B
  - 1luc_A
  - 1q6o_A
  - 1qop_A
  - 1qtw_A
  - 2v3g_A
  - 8a3h_A
- 52086
  - 2q3l_B
- 50813
  - 1zng_A
  - 1r0u_A
- 55247
  - 2p5m_B
- 46688
  - 2gfn_A
  - 1s3j_A
  - 2iu5_A
- 53334
  - 2fhp_B
  - 1g8a_A
  - 2ex4_B
- 53821
  - 1gud_B
  - 1usg_A
  - 3ckm_A
- 52008
- 50728

Sequence Alignment: Query Protein V.S Result Chain: 1gve_B

```
121 122 123 _ 178 179 180 _ 183 184 185 186 _ 197 198 199 200 201 _ 219 220 221 222 _ 229 230 231 232 233 234 _ 245 246 247 248 _ 256 257 258 259 260 261 262 263 264 265 266 _ 270 271 272 273 274 _ 284 285 286 _ 321 322 323 324
 M   R   P  _  R   A   N  _  G   Y   L   G  _  N   R   I   D   K  _  K   V   D   I  _  D   P   E   Y   L   Y  _  I   V   Y   A  _  S   V   R   G   I   A   G   M   R   K   A  _  G   V   V   V   M  _  V   P   P  _  Q   E   Y   F
 S   S   L  _  P   L   A  _  A   V   R   W  _  H   H   S   Q   L  _  V   D   L   F  _  L   K   P   A   D   V  _  F   Y   L   H  _  H   G   T   P   I   E   E   T   L   Q   A  _  G   W   I   M   P  _  V   S   W  _  F   G   L   R
286 287 288 _ 195 196 197 _ 263 264 265 266 _ 269 270 271 272 273 _ 103 104 105 106 _ 82  83  84  85  86  87  _ 106 107 108 109 _ 113 114 115 116 117 118 119 120 121 122 123 _ 157 158 159 160 161 _ 142 143 144 _ 186 187 188 189
 □   □   □     □   □   □     □   □   □   □     □   □   □   □   □     □   □   □   □     □   □   □   □   □   □     □   □   □   □     ☑   □   □   □   □   □   □   □   □   □   □     □   □   □   □   □     □   □   □     □   □   □   □
```

Query Protein Chain Length:	325
Result Protein Chain Length:	301
Alignment Length:	52
RMSD(unit:angstrom):	0.44862814677392304

Figure 3. ProteinDBS retrieval results visualization for local-to-local comparison. The top-left panel shows a superimposed view of a query protein and a selected result protein from the ranked list in the top-right panel (the first-ranked protein chain in second-ranked SCOP fold). Users can check 'Result Protein' and/or 'Query Protein' to view the retrieved chain and/or the query chain. By clicking on a hyperlink below the top-left panel, users can view different display themes. The lower panel shows the sequence alignment. Clicking on 'residue checkbox' below a pair of residue will highlight residues in the 3D superimposition view. To show one specific substructure, users first click hyperlink 'Clear Display' to hide all the substructures and then click substructure bar to show the designed substructure.

family, they are denoted as good results. As the global comparison method introduces new features to improve the system performance, on average, our system's global search exhibits 97.04% precision at the 10% recall rate and 87.82% precision at the 100% recall rate. A query using the protein chain with the maximum length in the testing set, 566 C_α atoms, takes 3.37 s to return the ranked search results. These tests were conducted on a Linux distributed system consisting of five servers (13).

We applied the local comparison method to SCOP fold classification and compared its performance with known algorithms, such as DALI (2), CE (3), MultiProt (18), SSM (19) and MAMMOTH (20), on a non-redundant database of 3276 protein chains selected from PDBSelect and SCOP v1.73. Our system was able to return ranked results in 182 s for query protein structures with an average length of 167 residues, which is 53.10, 10.87, 3.60 and 1.64 times faster than DALI, CE, MultiProt and MAMMOTH, respectively. Evaluated on three different data sets of non-redundant proteins from SCOP, the average accuracy of our system is approximately equal to DALI, better than MAMMOTH and significantly better than CE, MultiProt and SSM. These tests were conducted on a Linux Fedora server with AMD Opteron dual-core 1000 series processors and 2GB RAM (14).

DISCUSSION

Over the past decade, we have witnessed a rapidly increasing number of protein structures, which poses a great challenge to search engines that retrieve structurally similar proteins. The ProteinDBS v2.0 web server presented in this article comes equipped with an efficient and accurate search engine, a large-scale protein structure database and a more user-friendly interface. ProteinDBS can return accurate results in seconds for global structure search and takes much less time for local structural searches compared to other accurate comparison methods while preserving higher or similar accuracy. It is expected that this web server will be beneficial to the life sciences community for comparative structural analysis, automatic fold classification and the discovery of functional and evolutionary connections between protein structures.

ACKNOWLEDGEMENTS

The authors are grateful to the researchers and groups who made the following software packages and databases available for us to use in ProteinDBS: the Jmol package (http://jmol.sourceforge.net/), which generates view of aligned proteins, the SCOP database for ground truth testing (15), and the PDB (http://www.pdb.org) for maintaining the tertiary structures.

FUNDING

Shumaker Endowment in Bioinformatics. Funding for open access charge: University of Missouri Shumaker Endowment in Bioinformatics.

Conflict of interest statement. None declared.

REFERENCES

1. Zarembinski,T.I., Hung,L.W., Mueller-Dieckmann,H.J., Kim,K.K., Yokota,H., Kim,R. and Kim,S.H. (1998) Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics. *Proc. Natl Acad. Sci. USA*, **95**, 15189–15193.
2. Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.
3. Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineer.*, **11**, 739–747.
4. Martin,A.C. (2000) The ups and downs of protein topology; rapid comparison of protein structure. *Protein Engineer.*, **13**, 829–837.
5. Carpentier,M., Brouillet,S. and Pothier,J. (2005) YAKUSA: a fast structural database scanning method. *Proteins*, **61**, 137–151.
6. Yang,J.M. and Tung,C.H. (2006) Protein structure database search and evolutionary classification. *Nucleic Acids Res.*, **34**, 3646–3659.
7. Lo,W.C., Lee,C.Y., Lee,C.C. and Lyu,P.C. (2009) iSARST: an integrated SARST web server for rapid protein structural similarity searches. *Nucleic Acids Res.*, **37**, W545–W551.
8. Shyu,C.R., Chi,P.H., Scott,G. and Xu,D. (2004) ProteinDBS: a real-time retrieval system for protein structure comparison. *Nucleic Acids Res.*, **32**, W572–W575.
9. Leslie,M. (2004) DATABASE: Protein Matchmaking. *Science*, **305**, 1381b.
10. Friedberg,I. and Godzik,A. (2005) Connecting the protein structure universe by using sparse recurring fragments. *Structure*, **13**, 1213–1224.
11. Hvidsten,T.R., Laegreid,A., Kryshtafovych,A., Andersson,G., Fidelis,K. and Komorowski,J. (2009) A comprehensive analysis of the structure-function relationship in proteins based on local structure similarity. *PloS One*, **4**, e6266.
12. Soding,J. and Lupas,A.N. (2003) More than the sum of their parts: on the evolution of proteins from peptides. *BioEssays*, **25**, 837–846.
13. Chi,P.-H. (2007) Efficient protein tertiary structure retrievals and classifications using content based comparison algorithms. *Ph.D. Thesis*. University of Missouri-Columbia, Columbia, MO, USA.
14. Chi,P.H., Pang,B., Korkin,D. and Shyu,C.R. (2009) Efficient SCOP-fold classification and retrieval using index-based protein substructure alignments. *Bioinformatics*, **25**, 2559–2565.
15. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
16. Griep,S. and Hobohm,U. (2009) PDBselect 1992-2009 and PDBfilter-select. *Nucleic Acids Res.*, **38**, D318–D319.
17. Ciaccia,P., Patella,M. and Zezula,P. (1997) M-tree: an efficient access method for similarity search in metric spaces International Conference on Very Large Database, Athens, Greece, pp. 426–435.
18. Shatsky,M., Nussinov,R. and Wolfson,H.J. (2004) A method for simultaneous alignment of multiple protein structures. *Proteins*, **56**, 143–156.
19. Krissinel,E. and Henrick,K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr.*, **60**, 2256–2268.
20. Ortiz,A.R., Strauss,C.E. and Olmea,O. (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Prot. Sci.*, **11**, 2606–2621.