P3S: Protein Structure Similarity Search

Jakub Galgonek, Tomáš Skopal, and David Hoksza

Departement of Software Engineering, Charles University in Prague
Malostranske nám. 25, 118 00 Praha 1, Czech Republic
{galgonek,skopal,hoksza}@ksi.mff.cuni.cz

Abstract. Similarity search in protein structure databases is an important task of computational biology. To reduce the time required to search for similar structures, indexing techniques are being often introduced. However, as the indexing phase is computationally very expensive, it becomes useful only when a large number of searches are expected (so that the expensive indexing cost is amortized by cheaper search cost). This is a typical situation for a public similarity search service. In this article we introduce the P3S web application (`http://siret.cz/p3s`) allowing, given a query structure, to identify the set of the most similar structures in a database. The result set can be browsed interactively, including visual inspection of the structure superposition, or it can be downloaded as a zip archive. P3S employs the SProt similarity measure and an indexing technique based on the LAESA method, both introduced recently by our group. Together with the measure and the index, the method presents an effective and efficient tool for querying protein structure databases.

Keywords: protein structure, similarity retrieval, web service.

1 Introduction

Proteins have a wide range of functions in living organisms such as enzymatic, signaling, transportation and building function. The way proteins carry out their biological function is through interaction with other proteins or small molecules based on spatial arrangement of their physicochemical properties. Therefore, identification of similar tertiary folds can bring invaluable insight into function of proteins with known 3D structures [1]. The similarity methods can be divided into two groups — global similarity measures taking into account the whole structure, and local similarity measures comparing only active sites, that is, the sites where the protein links to its binding ligand or protein. The latter methods can be more precise in linking functional similarity to structural similarity since they focus directly on the site where the action occurs. On the other hand, for new structures the active site is often not known as localizing the exact ligand position requires the structures to be resolved at high resolution. In the current version (v.2011) of sc-PDB [2], the annotated database of druggable binding sites from the Protein DataBank (PDB) [3] contains binding site information for 3034 proteins and, at the same time (March 27, 2012), PDB contains 3D information for 80,402 structures. This disproportion favors the first group of

I. Caragiannis et al. (Eds.): Euro-Par 2012 Workshops, LNCS 7640, pp. 228–237, 2013.

methods which are able to assess pairwise protein structure similarity without the need of knowing exact binding site positions. There have been many methods proposed for assessing protein similarity on the tertiary structure level in recent years [4–9]. However, one can see the trend of addressing not only the effectiveness (quality) of the comparison process but also its efficiency (speed). The shift can be clearly attributed to the growth of the number of structures in PDB. On the other hand, we stress out that the need for development of fast methods is still not saturated if we consider the difference in the number of structures deposited in sequence and structure databases. Currently (March 27, 2012), PDB contains 80,402 structures in comparison to 535,248 sequences deposited in UniProtKB/Swiss-Prot [1]. Since it has been shown that the structure is more conserved than sequence [10] there is a strong demand on increasing the quality and speed of structure determination techniques. In turn, the demand results in more structures with the ultimate goal to fully cover the sequence space by structural analogues.

Recently, we have contributed to the global protein structure similarity area by introducing a method called SProt [11]. SProt not only focuses on high-quality structure alignment but also introduces database indexing techniques applicable to the proposed measure, thus greatly improves the efficiency when used for searching large protein structure databases.

To further address the needs of bioinformatics community, the methods should provide not only similarity search of user-uploaded structures in protein databases but also subsequent visual inspection of the superposition. That can further help in explanation of the functional relation between the query and the results. Although there exist some exceptions (see section 3.1), these requirements are the weak points of many other methods. In this article, we present a freely available interactive web application for similarity search in protein structure databases allowing querying a database of protein structures using a single query structure. It utilizes the previously presented SProt [11] method which proved to be a highly effective and efficient addition to the portfolio of protein structure similarity methods. The application allows to visually inspect the alignments and the superpositions of the results with the query structure, while it also provides an export of the results for later use.

2 Implementation

In the following section we briefly summarize the technologies that have been used to implement our system. Subsequently, we focus on the description of the usage of the web application.

2.1 Web Server

From the beginning, the P3S web server has been considered as an interactive web application. Because of this intent, we use Google Web Toolkit[1] that has

[1] `https://developers.google.com/web-toolkit`

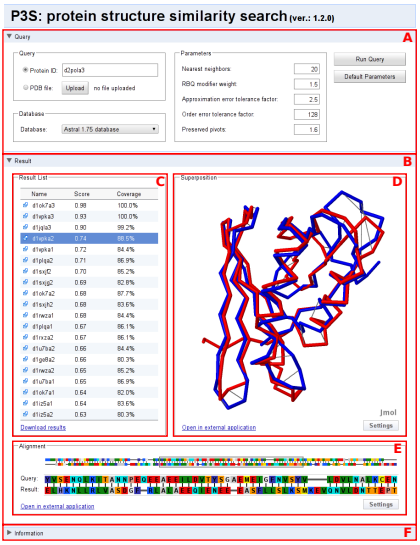

Fig. 1. The P3S web server user interface

been designed especially for writing such type of applications. The framework allows to develop the client-side part of the application (running in a user's web browser) together with the required application server. The Jmol java applet[2] has been used for the dynamic visualization of the proteins.

The computational server employs the SProt similarity measure and an indexing technique based on the LAESA method (adjusted specifically for the SProt measure), both introduced recently by our group [11]. In order to obtain a high-performance solution, the computational server has been developed in C++ using *Intel Threading Building Blocks* (TBB) library [12] that provides parallel implementation. The computational server is running on a dedicated server and communicates with the application server using the well-established CORBA technology.

2.2 Usage of P3S

The web server is available at the address `http://siret.cz/p3s`. The user interface is divided into three panels — Query (Fig. 1A), Result (Fig. 1B) and Information (Fig. 1F). The Query panel is used to submit a query. The Result panel is used to present the retrieved structures and the Information panel shows link to the application documentation and other relevant information.

Query Submission. To perform a search, the user has to select a query structure, the target database, and the number of the most similar structures which are to be retrieved from the database.

[2] `http://www.jmol.org`

The query structure can be defined by its ID or uploaded manually. In the first case the structure stored on the server having the given code is used. The second choice is to upload user-defined PDB file. The PDB file should use the actual version of the PDB format and should also include all heavy atoms. The application allows checking the messages from the PDB parser, which is useful in cases when the PDB file is rejected due to errors in the format. If the PDB file contains multiple models, only the first one is taken into account.

Currently, P3S supports two databases that the user can choose. The first one is the Astral [13] 1.75 database containing 33,352 structures (only structures having different sequences are included). The second one is the ProtDex2 database containing 34,055 structures [9]. It is applicable mainly for comparison with other methods. Note that the databases do not contain multi-domain protein structures. Therefore the query should be preferably a single-domain structure.

The number of the most similar structures to be retrieved can be set together with other parameters that influence the behavior of the access method. The most important parameter is so-called *RBQ modifier weight*, which controls the efficiency vs. precision tradeoff.

Although setting the number of retrieved structures and the RBQ modifier weight can speedup the search substantially, the time needed for the evaluation of a query depends also on the size of the query structure and on the number of (similar) structures in the database. To give the user information on the query progress, the query progress bar shows the fraction of the already searched database. Note that because of the non-linear behavior of the search algorithm the progress pace may not be constant during the querying.

Result Presentation. When the query process completes, the retrieved structures are presented in the result list (Fig. 1C) containing several columns. The *Name* column describes the SCOP ID [14] of the result structure. The *Score* column shows similarity score between the query structure and given result structure. The score range goes from 0 to 1. The last column, *Coverage*, determines the coverage of the query structure, i.e., percentage of the amino acids of the query that were aligned with some amino acid of the given result structure. The leftmost column contains links to the PDB server, showing detailed information about the selected structure. Structures are sorted according to the *Score* column.

When selecting an item from the result list, visualizations of the superposition (Fig. 1D) and the alignment (Fig. 1E) between the query structure and the selected structure are shown. The orientation of the query structure is not changed when switching between the items. Hence, if the user focuses on some part of the query structure, switching to other result item does not distract her/him.

The alignment and superposition visualizations are interlinked so that one can inspect the alignment on both sequence and structure level. In the superposition visualization, the lines between the C_α atom positions indicate the aligned residues. Furthermore, when the mouse hovers over an amino acid in the alignment visualization, the corresponding amino acid is highlighted in the

superposition visualization and vice versa. Both of the visualizations can be customized using the *Settings* menu placed in the lower right corner.

Results Download. The results can be also exported in the form of a zip archive which allows a possible later analysis of the obtained results. The archive includes the query PDB file, a text file describing the query parameters and a text file containing the brief descriptions (name, score and coverage) of the obtained result set. For each retrieved structure, the zip archive contains subfolder with other useful information related to the retrieved structure — the original PDB file, the superimposed PDB file, the superposition (in the Jmol format), the alignment (in the FASTA format), and the information file containing additional information about the retrieved structure, especially the transformation defining the superposition.

If the user does not want to use the Jmol for the visualization/analysis of the superposition, (s)he has two possibilities. The user can just enter the PDB file with query structure and the PDB file with superimposed result structure into a program of her/his choice. Another possibility is to use the original PDB file and apply the transformation stored in the information file.

3 Results

In this section, we focus on a comparison of our web application with other current web applications used for similarity search in protein structure databases. To evaluate the effectiveness of P3S we compare it to other, not necessarily web-based, methods.

3.1 Comparison of Web Application Interfaces

For our first comparison, we have selected applications which we think represent the current state of the art in the field, namely: Vorometric[3] [5], ProBiS[4] [15], deconSTRUCT[5] [16], PDBeFold[6] [17], VAST[7] [18], iSARST[8] [19], 3D-BLAST[9] [20], and Dali server[10] [21]. The summary of basic features of the application is shown in Table 1.

Each of the compared web applications allows to select a query structure by its SCOP ID (see row *support SCOP ID*) or by its PDB ID (see row *support PDB ID*). However, much more important is the ability to upload a user's protein structure. This feature is supported by all the applications except Vorometric

[3] `http://bio.cse.ohio-state.edu/Vorometric/`
[4] `http://probis.cmm.ki.si/`
[5] `http://epsf.bmad.bii.a-star.edu.sg/struct_server.html`
[6] `http://www.ebi.ac.uk/msd-srv/ssm`
[7] `http://www.ncbi.nlm.nih.gov/Structure/VAST/`
[8] `http://sarst.life.nthu.edu.tw/iSARST/`
[9] `http://3d-blast.life.nctu.edu.tw/`
[10] `http://ekhidna.biocenter.helsinki.fi/dali_server/`

Table 1. Basic features of web applications

	P3S	Vorometric	ProBiS	deconSTRUCT	PDBeFold	VAST	iSARST	3D-BLAST	Dali server
upload PDB file	✓	–	✓	✓	✓	✓	✓	✓	✓
support SCOP ID	✓	✓	–	–	✓	–	✓	✓	–
support PDB ID	✓	✓	✓	✓	✓	✓	✓	✓	✓
email notification	–	✓	✓	✓	–	–	–	–	✓
search history	–	✓	–	–	–	–	✓	–	–
presentation scheme	C	A	C	A	A	B	A	A	B
external links	✓	✓	✓	✓	✓	✓	✓	✓	–
visual inspection	✓	✓	✓	✓	✓	–[1]	✓	✓	✓
superposition download	✓	✓	✓	✓	✓	✓	✓	✓	✓
alignment download	✓	✓	–	✓	✓	✓	–	–	–
full download (zip)	✓	✓	–[2]	✓	–[2]	–[2]	–	–[2]	–[2]

[1] Superpositions can be visualized by an external application.
[2] It is possible to download a text-based result list.

(see row *upload PDB file*). Some of the applications also allow to specify an email address to which a notification will be sent when the search is finished (see row *email notification*). It is very useful mainly for applications where searching takes a very long time.

Two of the applications also support searching history (see row *search history*) that allows to show results of older queries. In Vorometric, the history is showed after entering an email address. We think that it is not an optimal solution as the search history of a person can be displayed to anyone who knows the email address. The history of iSARST is based on the web browser session, so its usage is limited.

In the scheme of the presentation of retrieved structures, we can observe three basic types (see row *presentation scheme*). The result of searching is often presented as a list of the retrieved structures, while a separate page (type A) or separate pages (type B) are used to display the result details. The separate pages allow to display more details, however, the browsing through retrieved structures is not so comfortable. Instead, P3S (and ProBiS) display the result list together with the details about a selected structure (type C) on a single page.

The most important part of each application is the presentation of results itself — it includes the visualization of the alignment and the visualization of the superposition between a query structure and a retrieved structure. Except VAST that uses an external application, all applications display superpositions as an integral part of the web application (see row *visual inspection*). For this purpose, the Jmol java applet is used. Some of the applications use the superposition visualization also to show other details, e.g., to highlight aligned parts. However, the connection between the alignment visualization and the superposition visualization is a unique feature of P3S. Except Dali server, the other applications also present a link to other external resources (for example into PDB or SCOP database).

Another important part of the applications is the possibility to download the results. All applications allow to download the superposition of a query and a selected result structure (see row *superposition download*). On the other hand, the options to download the structure-based alignment for the select result structure are quite limited (see row *alignment download*). Also a possibility to download all results as a single file archive is not often supported, although many applications have the possibility to download text files with the list of the retrieved structures (see row *full download (zip)*).

Based on the above comparison we consider the interface of P3S to be comparable with interfaces of other web applications.

3.2 Comparison of Databases

Next, we focus on the supported databases and the query times. The results are summarized in Table 2.

To obtain a result in a reasonable time, large databases are generally employed by applications that use fast algorithms and simpler models. They are also employed by applications that do not present the result instantly. The applications based on slower algorithms using more precise models employ smaller but representative structure databases. P3S belongs to the second category. The supported databases are present in column *Database*, the size of the largest supported databases is presented in column *Max. size*.

Due to the differences in the databases, due to the hardware and also due to differences in weakness points of the respective algorithms, it is problematic to compare a query time of the applications. The query time is present in column *Query time* and represents a time declared by the respective authors wherever possible. In other cases, we present an approximate time based on our observation. Although these times cannot be used for rigorous comparison, they reflect the user experience time very well.

Although P3S uses an algorithm based on a precise model, the achieved query times are reasonably low to be considered as instant.

3.3 Effectiveness of P3S

Qualities of the SProt measure and its indexing method have been evaluated in the original paper describing the method [11]. In this new article, we want to introduce a modified experiment that evaluates the overall quality of the measure and its indexing method.

For this purpose we adopted the experimental setup from the area of *information retrieval*. Such a setup comprises of a database of objects that a tested system uses for retrieval and a query set of objects. For each object from the query set, only a (small) subset of the database is considered to be relevant by the author of the experiment. When the information system returns the result for the given query, the quality of such result can be described in terms of *precision* and *recall*. Precision expresses how many percent of objects in the result set are relevant. Recall expresses how many percent of all relevant objects are obtained

Table 2. Used databases and query times

	Database	Max. size	Query time
P3S	SCOP/ASTRAL-100 (ver. 1.75), ProtDex2 dataset (SCOP/ASTRAL)	34,055	The search takes seconds to minutes.
Vorometric	SCOP/ASTRAL-25 (ver. 1.73)	6,981	The search may take a few minutes.
ProBiS	PDB (non-redundant structures)	30,300	Computation for a medium sized protein will take around 30-60min.
deconSTRUCT	PDB chains + representative subsets	179,543	Based on our observation, the search takes seconds to minutes.
PDBeFold	PDB chains, SCOP/ASTRAL (ver. 1.73) + subset	192,994	Most requests are completed within 1 or 2 minutes.
VAST	PDB + medium-redundancy subset	N/A	Based on our observation, the search takes minutes to tens of minutes.
iSARST	PDB chains (May 2011) + subsets, SCOP/ASTRAL (ver. 1.73) + subsets	176,690	A typical search along with superimposing 100 structures takes only 3∼5 seconds.
3D-BLAST	PDB chains (03-Mar-12) + nr subset, SCOP/ASTRAL (ver. 1.75) + subsets	173,232	The method search more than 10000 protein structures in 1.3 seconds.
Dali server	PDB chains	190,642	Most queries finish within an hour.

in the result set. These two qualities usually tend to be in a contradiction. An effort that increases the precision decreases the recall, and vice versa.

To evaluate the dependency between precision and recall for an average query, the whole query set instead of just one query is used. For each of the query from the query set, the precision at the given recall level is computed and averaged across the whole query set. This is performed for different values of recall plotted in the precision-recall graph.

In our experiment, we used the ProtDex2 dataset consisting of 34,055 proteins. As the query set, 108 structures from medium-size families of the dataset were selected [9]. The retrieved structure is considered to be relevant to the query if it comes from the same SCOP family [14]. This dataset is widely used, so a comparison with other methods can be performed. For the comparison, we have selected the following methods: FAST [4], Vorometric [5], CE [6], SARST [7], 3D-BLAST, [8], PSI-BLAST [22], and ProtDex2 [9]. The graph data for the compared methods are borrowed from [5] and [7].

We tested P3S for different values of the weight of the RBQ modifier, while other parameters remained unchanged. The reason for such decision was the weight had the greatest impact on the effectiveness and efficiency, because it was the only parameter directly affecting the measure and its indexability.

Figure 2 shows the results of the experiment. With increasing weight w, the precision of the method is decreasing, especially for high recall levels. Because for $w > 2.0$ the impact on speed is negligible, we strongly recommend using the weight in the range $\langle 0, 2.5 \rangle$. As the figure shows, P3S has better precision-recall curve than the other methods, except Vorometric and FAST. In comparison with Vorometric, the curves of P3S are the same or slightly worse for medium recall levels, while they are noticeably better for high recall levels. FAST slightly outperforms P3S but on the other hand it does not offer a web search server as P3S does and therefore is not easily accessible to wide scientific community which makes P3S one of the most precise publicly available web application in the dataset.

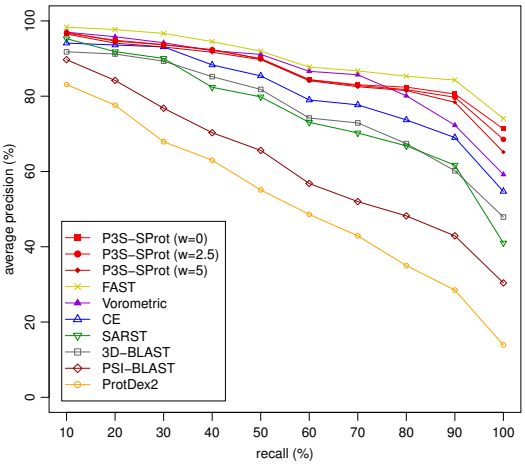

Fig. 2. Average precision-recall curves

4 Conclusion

In this paper, we have introduced a freely available web application P3S designed for similarity search in protein structure databases. The application obtains a protein structure as its input and returns the set of most similar structures. The result set can be browsed interactively or it can be downloaded as a zip archive. The application employs the SProt measure and an access method, both developed in our research group. The measure and the access method achieve very good results in comparison with other methods as was shown.

Acknowledgement. This work was supported by the Grant Agency of Charles University [project Nr. 430711]; the Czech Science Foundation [project Nr. 202/11/0968]; the Federation of European Biochemical Societies [Short-Term Fellowship]; and the Specific Academic Research [project Nr. SVV-2012-265312].

References

1. Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., Thornton, J.M.: CATH–a hierarchic classification of protein domain structures. Structure (London, England: 1993) 5(8), 1093–1108 (1997)
2. Meslamani, J., Rognan, D., Kellenberger, E.: sc-PDB: a database for identifying variations and multiplicity of 'druggable' binding sites in proteins. Bioinformatics 27(9), 1324–1326 (2011)
3. Berman, H.M., Westbrook, J.D., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The Protein Data Bank. Nucleic Acids Res. 28(1), 235–242 (2000)

4. Zhu, J., Weng, Z.: FAST: a novel protein structure alignment algorithm. Proteins 58(3), 618–627 (2005)
5. Sacan, A., Toroslu, H.I., Ferhatosmanoglu, H.: Integrated search and alignment of protein structures. Bioinformatics 24(24), 2872–2879 (2008)
6. Shindyalov, I.N., Bourne, P.E.: Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng. 11(9), 739–747 (1998)
7. Lo, W.C., Huang, P.J., Chang, C.H., Lyu, P.C.: Protein structural similarity search by Ramachandran codes. BMC Bioinformatics 8 (2007)
8. Tung, C.H.H., Huang, J.W.W., Yang, J.M.M.: Kappa-alpha plot derived structural alphabet and BLOSUM-like substitution matrix for rapid search of protein structure database. Genome Biol. 8(3), R31 (2007)
9. Aung, Z., Tan, K.L.: Rapid 3D protein structure database searching using information retrieval techniques. Bioinformatics 20(7), 1045–1052 (2004)
10. Chothia, C., Lesk, A.M.: The relation between the divergence of sequence and structure in proteins. The EMBO Journal 5(4), 823–826 (1986)
11. Galgonek, J., Hoksza, D., Skopal, T.: SProt: sphere-based protein structure similarity algorithm. BMC Proteome Science 9(suppl. 1) S20 (2011)
12. Reinders, J.: Intel threading building blocks: outfitting C++ for multi-core processor parallelism. O'Reilly Media, Inc. (2007)
13. Chandonia, J.M.M., Hon, G., Walker, N.S., Lo Conte, L., Koehl, P., Levitt, M., Brenner, S.E.: The ASTRAL Compendium in 2004. Nucleic Acids Res. 32(Database issue), D189–D192 (2004)
14. Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C.: SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247(4), 536–540 (1995)
15. Konc, J., Janezic, D.: ProBiS: a web server for detection of structurally similar protein binding sites. Nucleic Acids Res. 38(Web-Server-Issue), 436–440 (2010)
16. Zhang, Z.H., Bharatham, K., Sherman, W.A., Mihalek, I.: deconSTRUCT: general purpose protein database search on the substructure level. Nucleic Acids Res. 38(Web-Server-Issue), 590–594 (2010)
17. Krissinel, E., Henrick, K.: Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallographica Section D 60(12 pt. 1), 2256–2268 (2004)
18. Gibrat, J.F., Madej, T., Bryant, S.H.: Surprising similarities in structure comparison. Current Opinion in Structural Biology 6(3), 377–385 (1996)
19. Lo, W.C., Lee, C.Y., Lee, C.C., Lyu, P.C.: iSARST: an integrated SARST web server for rapid protein structural similarity searches. Nucleic Acids Research 37(Web-Server-Issue), 545–551 (2009)
20. Yang, J.M.M., Tung, C.H.H.: Protein structure database search and evolutionary classification. Nucleic Acids Research 34(13), 3646–3659 (2006)
21. Holm, L., Rosenström, P.: Dali server: conservation mapping in 3D. Nucleic Acids Research 38(Web-Server-Issue), 545–549 (2010)
22. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25(17), 3389–3402 (1997)