

Protein 3D Structure and Electron Microscopy Map Retrieval Using 3D-SURFER2.0 and EM-SURFER

UNIT 3.14

Xusi Han,¹ Qing Wei,² and Daisuke Kihara^{1,2}

¹Department of Biological Sciences, Purdue University, West Lafayette, Indiana

²Department of Computer Science, Purdue University, West Lafayette, Indiana

With the rapid growth in the number of solved protein structures stored in the Protein Data Bank (PDB) and the Electron Microscopy Data Bank (EMDB), it is essential to develop tools to perform real-time structure similarity searches against the entire structure database. Since conventional structure alignment methods need to sample different orientations of proteins in the three-dimensional space, they are time consuming and unsuitable for rapid, real-time database searches. To this end, we have developed 3D-SURFER and EM-SURFER, which utilize 3D Zernike descriptors (3DZD) to conduct high-throughput protein structure comparison, visualization, and analysis. Taking an atomic structure or an electron microscopy map of a protein or a protein complex as input, the 3DZD of a query protein is computed and compared with the 3DZD of all other proteins in PDB or EMDB. In addition, local geometrical characteristics of a query protein can be analyzed using VisGrid and LIGSITE^{CSC} in 3D-SURFER. This article describes how to use 3D-SURFER and EM-SURFER to carry out protein surface shape similarity searches, local geometric feature analysis, and interpretation of the search results. © 2017 by John Wiley & Sons, Inc.

Keywords: 3D Zernike descriptors • electron microscopy map • local geometric feature • protein structure comparison • protein surface comparison

How to cite this article:

Han, X., Wei, Q., & Kihara, D. (2017). Protein 3D structure and electron microscopy map retrieval using 3D-SURFER2.0 and EM-SURFER. *Current Protocols in Bioinformatics*, 60, 3.14.1–3.14.15. doi: 10.1002/cpbi.37

INTRODUCTION

Proteins perform a vast array of functions and participate in essentially every cellular process. The tertiary structure of proteins provides a physical platform for carrying out functions. The structure information of proteins thus forms the basis for understanding the principles of life and aids in developing new strategies to regulate biological pathways and other processes. Since the protein structure is directly related to their molecular function, similarity in the structures is more preserved than similarity at the sequence level (Wilson, Kreychman, & Gerstein, 2000). Therefore, structure-based protein comparison is capable of revealing remote relationships that are hard to detect from sequences.

With the exponential growth in the number of solved protein structures in the Protein Data Bank [PDB; Berman et al., 2000; also see *UNIT 1.9* (Di Costanzo, Ghosh, Zardecki, & Burley, 2016)] and Electron Microscopy Data Bank (EMDB; Lawson et al., 2016), it has become crucial to develop search tools that can compare protein structures and help us understand the relationship between them. The conventional approach for protein

**Finding
Similarities and
Inferring
Homologies**

3.14.1



Current Protocols in Bioinformatics 3.14.1–3.14.15, December 2017
Published online December 2017 in Wiley Online Library (wileyonlinelibrary.com).
doi: 10.1002/cpbi.37
Copyright © 2017 John Wiley & Sons, Inc.

Supplement 60

structure comparison is aligning atoms or residues of proteins. This approach is time-consuming, as sampling of different orientations in the three-dimensional (3D) space is needed, making it inappropriate for searching against a whole structure database in real time. The use of 3D Zernike descriptors (3DZD; Kihara, Sael, Chikhi, & Esquivel-Rodriguez, 2011) was proven to be suitable for efficient structure comparisons in previous work (Esquivel-Rodriguez et al., 2015; La et al., 2009; Sael et al., 2008; Xiong, Esquivel-Rodriguez, Sael, & Kihara, 2014). Such descriptors represent a 3D object in a compact vector and in a rotation-invariant fashion, enabling fast search against structure databases.

We have developed 3D-SURFER (La et al., 2009; Xiong et al., 2014; <http://kiharalab.org/3d-surfer/index.php>) and EM-SURER (Esquivel-Rodriguez et al., 2015; <http://kiharalab.org/em-surfer/index.php>), which are Web-based platforms for high-throughput protein structure comparison and analysis. Structures in each server are automatically synchronized with PDB or EMDB weekly. Currently, 3D-SURFER holds over 500,000 entries (including chain, domain, and complex structures) and EM-SURER holds over 3,000 maps at various map resolutions. Both of them utilize 3DZD to extract global surface information from proteins and quantify shape similarity by calculating Euclidean distance between a pair of 3DZDs. In 3D-SURFER, the VisGrid (Li et al., 2008) and LIGSITE^{CSC} (Huang & Schroeder, 2006) algorithms are employed to characterize local geometric features of a query protein, including pocket, cavity, protrusion, and flat regions.

BASIC PROTOCOL 1

SEARCHING PROTEIN 3D STRUCTURES USING 3D-SURFER

The input required to run the online 3D-SURFER server is either the structure ID in PDB or a PDB-format atom-coordinate file. Top retrieved structures and their information, as well as local analysis of a query protein, are presented in a results Web page.

Necessary Resources

Hardware

Any up-to-date computer with Internet access

Software

A Javascript-enabled Web browser, such as Internet Explorer, Firefox, or Chrome

Files

Besides using the search box to enter a specific structure identification (ID) code from PDB, users can also upload their own structure file. The file to upload should be in PDB format with atom coordinates. If analysis of a structure domain is desired, the amino acid range of the domain needs to be specified in the domain range box. The whole structure is used to run the search if no domain range is specified.

1. Open the browser and go to the URL <http://kiharalab.org/3d-surfer/index.php>. Identify the Search button in the top panel, and click it to go to the search page. Figure 3.14.1 illustrates the search page of 3D-SURFER.
2. To enter a query protein, users can either type the PDB ID of the protein or upload a PDB-format file. The structure ID box accepts three categories of inputs: chain, domain, and complex. When entering structure ID in the box, structures in all three categories having the same ID as entered will appear in a drop-down menu, which can be scrolled down and selected.

A specific chain in a PDB structure can be entered as 1h41-B, which is PDB ID followed by a hyphen and the chain ID. A domain of a chain can be specified by further adding

Submit a protein

Please refer to the tips below when uploading a file:

- It is possible that the structure ID already exists in the database. Try to use the search box first
- If you benchmark our program, please [access the page](#)

Step 1 (Query protein)

Structure ID: <input type="text" value="3qd8-A"/> e.g. Chain ID: 7tim-A Complex ID: 2wiw or 12e8-C01 Domain ID: 1h41-B-02	Or	Upload a structure file: <input type="button" value="Choose File"/> no file selected An example file you can upload. (Optional) Please specify your domain range in your uploaded file: <input type="text"/>
--	----	--

Step 2 (Representation)

Surface representation:	<input checked="" type="radio"/> All atom <input type="radio"/> Main chain atom
-------------------------	---

Step 3 (Database)

Template database:	Chain <input type="button" value="v"/>
--------------------	--

Step 4 (Filter)

CATH filter:	None <input type="button" value="v"/>
Length filter:	<input checked="" type="radio"/> ON <input type="radio"/> OFF

Figure 3.14.1 Screenshot of the job submission page in 3D-SURFER.

domain ID coming from the CATH annotation (e.g., 1h41-B-02). A complex is entered as 1h41 or 12e8-C01, depending on the number of complexes identified in the corresponding PDB structure. Complexes within a PDB structure are identified by calculating the number of interacting atoms in each pair of chains using 4.5 Å as cutoff. After entering a complex ID, the chain composition of the complex entered will appear below the structure ID box.

- Select a surface representation method. All atom representations utilize coordinates of all atoms in the structure to build the global surface. The main chain atom representation only includes C α , C, and N atoms in the main chain. The choice should be made according to the purpose of the search. In general, if the purpose is to find structures with the same fold classification as CATH (Sillitoe et al., 2015; see *UNIT 1.28*) or SCOP [Andreeva et al., 2008; also see *UNIT 1.26* (Andreeva, Howorth, Chothia, Kulesha, & Murzin, 2015)], the main chain atom representation performs best as long as the query molecule has a globular shape; the all-atom representation performs best if a query has a long tail or unstructured loop region. For the benchmark studies, see our previous work (Sael & Kihara, 2010).
- Select the structure database to search. Similar to selecting input structure categories, there are chain, domain, and complex template databases, and also a database that contains all three of them.
- Specify the CATH filter and the length filter. To filter out similar structures, users can specify up to which CATH hierarchy level they want to use in the search. By

default, no CATH filter is applied to the search. If the length filter is enabled, only proteins whose sizes are between 0.57 and 1.75 times the size of the query protein will be retrieved in the results page. The length filter is on by default. The length filter is useful if users want to retrieve proteins with a similar size to the query protein. Since 3DZD considers shape similarity but ignores size, proteins of a similar shape but different size can be retrieved with the length filter off.

There are all four levels of CATH hierarchy: class (C), architecture (A), topology (T), and homologous superfamilies (H) (Sillitoe et al., 2015). Five types of CATH filters are provided: 'C', 'CA', 'CAT', 'CATH', and 'None'. For example, if the 'CA' filter is used, only one structure with a certain class and architecture code combination will appear in the results page. All other structures with the same CA code combination will be filtered out.

6. Click the Submit button to submit the job. The Reset button restores all the settings to default.

To submit a batch of queries, click on the Benchmark button on the top panel. Users can either type a list of structure IDs or upload a structure ID list file. Most options are the same as those introduced before. One difference is that users can pick how many top results they want to obtain in the output file using the scroll-down window in step 5. The result for each query is listed in a separate file. Results for all queries submitted can also be downloaded as a single compressed zip file.

7. Users can also start a search in 3D-SURFER by specifying a URL, for example, <http://kiharalab.org/3d-surfer/cgi-bin/search.php?q=7tim&atomtype=cacn&database=chain>. Instead of entering a query ID on the 3D-SURFER search page, users can type the Web link in the URL bar directly and press Enter to start the search. This URL points to the 3D-SURFER results page for the query and filters specified. As 3D-SURFER does not assign a job ID for each search submission, users need to start the search again if they would like to see the search result using settings they picked previously. This URL gives users a quick access to the 3D-SURFER result page, without picking parameters on the search page. Also, users can link the structure of interest to the 3D-SURFER result page directly using this URL.

In the URL, query PDB ID is specified following q=, which accepts all three structure categories. Users can also specify surface presentation and template databases to use in the search (optional). The default filter setting (atomtype=all and database=chain) is employed if not provided in the link. Surface representation is specified by adding &atomtype= to the link following the query ID. all is for all atom representation and cacn is for main chain atom representation. The database to search against is specified by adding &database= to the link. chain, domain, complex, and all, respectively, represent chain, domain, complex, and all three databases together. The order of the surface representation method and template database is interchangeable in the link. In the example shown above, we use complex 7tim as query, main chain atom representation, and chain database in the search.

BASIC PROTOCOL 2

SEARCHING ELECTRON MICROSCOPY MAPS USING EM-SURFER

To start a search in EM-SURFER, the only input required is either the four-digit EMDB entry ID or the electron microscopy map for the query structure. Retrieved similar structures are presented in the form of a Web page. The steps for submitting a query to the EM-SURFER server are described below.

Necessary Resources

Hardware

Any up-to-date computer with Internet access

Software

A Javascript-enabled Web browser, such as Internet Explorer, Firefox, and Chrome

Submit an EM map

Please refer to the tips below when uploading a file:

- It is possible that the EMDB ID already exists in the database. Try to use the search box first.
- If you benchmark our program, please [access the page](#).

(For the purpose of review, please directly click the **Submit** button to get the search result, with the default options provided.)

Step 1 (Representation)

Contour shape representation:

EMDB contour

Step 2 (Query entry)

Enter 4-digit EMDB entry ID:
e.g., ID: 1884

Or

Upload an EM map (.map or .mrc) file ([Upload troubleshooting](#)):

Choose File no file selected [An example file.](#)

Recommended contour level: e.g., 3.16

Step 3 (Filter)

Volume filter:
(The volume of the EM entry in the database is between 0.8 and 1.2 times the volume of the query entry if ON, or else if OFF)

☒ ON ☐ OFF

Resolution filter:
(The query is only compared against maps in this resolution range. If both are left blank, no filtering is applied. If only one is provided, it will be the only restriction imposed)

Min: Max:

Submit

Reset

Figure 3.14.2 Screenshot of the job submission page in EM-SURFER.

Files

To enter a query structure, users can either provide an entry ID in EMDB or upload an electron microscopy map in MAP or MRC format. Users should specify the recommended contour level for the uploaded map file to use in construction of 3D map shape. The search will fail if no contour-level information is provided.

1. Open a Web browser and go to the URL <http://kiharalab.org/em-surfer/index.php>. Click the Search button at the top panel to open the search page (Fig. 3.14.2).
2. Specify the contour level that is used to represent the 3D shape of the query map. There are five options provided: EMDB contour, EMDB contour + 1/3 core, EMDB contour + 2/3 core, EMDB contour + 1/3 core + 2/3 core, and EMDB contour + 1 std.

“EMDB contour” is the recommended contour level suggested by the author of the query structure in EMDB. It is the default setting if no other is specified. “EMDB contour + 1/3 core” and “EMDB contour + 2/3 core” combine the 3DZD generated using “EMDB contour” (121 invariants) and the 3DZD generated using 1/3(max density) or 2/3 *(max density) as contour level (121 invariants). Those are the concatenations of two sets of 3DZD (242 invariants), and are able to represent regions closer to the core of the molecule. “EMDB contour + 1/3 core + 2/3 core” is a combination of three sets of 3DZD (363 invariants) that captures information at different depths of the structure. “EMDB contour + 1 std” is a concatenation of the “EMDB contour” and 3DZD generated from the isosurface at one standard deviation (242 invariants).*

3. Provide the query information. Users can provide the four-digit EMDB entry ID or upload an EM map. When the upload option is used, the input map should be in the MAP or MRC format. Users should also specify the contour level they want to use.
4. Choose the volume and the resolution filters. As size information of a protein is not reflected in its 3DZD descriptor, users can enable the volume filter to retrieve entries with similar volume (0.8 to 1.2 times the volume of the query). If users want to retrieve only structures within a certain resolution range, they can specify the resolution range in the resolution filter. If only a maximum or a minimum value is entered, it will be the only resolution restriction imposed. No resolution filter is applied by default.
5. Click the Submit button to start the search.

Similar to 3D-SURFER, if users want to submit a batch of entries, they can utilize the batch mode of EM-SURFER by clicking the Benchmark button at the top panel. Users can either type in all structure IDs or upload a list file. There is an option to specify the number of top results shown in the final output. By choosing from 10, 20, or 30 in the drop-down menu, users will get a corresponding number of retrieved structures for each query. The output for each query submitted can be downloaded separately or as a single compressed zip file.

GUIDELINES FOR UNDERSTANDING RESULTS

3D-SURFER Results

The results page shows the top 25 structures that share similar global surface shape to the query protein. We use the Euclidean distance between the 3D Zernike descriptors (3DZD) of a pair of proteins to quantify their similarity. Details of 3DZD calculation are introduced in the Background Information section, below. Empirically speaking, the surface similarity between two proteins is significant if the distance is below 10. Besides surface comparison results, users can also analyze geometric features of the query protein and run structure alignment between the query protein and a specific retrieved structure. Below, we explain and discuss the search results using 3qd8-A as an example, which is chain A in ferritin BrfB from *Mycobacterium tuberculosis* (Khare et al., 2011).

Local Features of Query Protein

At the top of the results page are shown the query ID, filters enabled in the search, the length of the query protein, and its CATH ID if available (Fig. 3.14.3). The visualization of the query protein is generated by the Jsmol applet at the top left panel. The representation of the protein can be changed using the Jsmol menu by clicking the right button of the mouse. To analyze geometric features of the query protein, click the Cavity, Protrusion, or Flat button to color residues that correspond to specified geometry on the protein surface calculated by VisGrid (Li et al., 2008). Red means the largest cavity or protrusion, green means the second largest, and blue means the third largest. Flat regions are colored in yellow.

The VisGrid algorithm identifies local geometric features of protein surfaces using the visibility criterion. Visibility is defined as the fraction of open directions from a target position on the protein surface. Thus, a protrusion is defined as a region that has high visibility while a cavity is a region with low visibility.

The surface area and volume for each cavity, protrusion, or flat region are calculated for the convex hull formed by all residues in that region. Convex hull is the smallest polygon that contains all the residues in that region. It is computed with the Qhull program (Barber, Dobkind, & Huhdanpaa, 1996).

Identification of surface pockets by LIGSITE^{CSC} (Huang & Schroeder, 2006) is invoked by clicking the Pocket button. If identified, the first, second, and third largest pocket

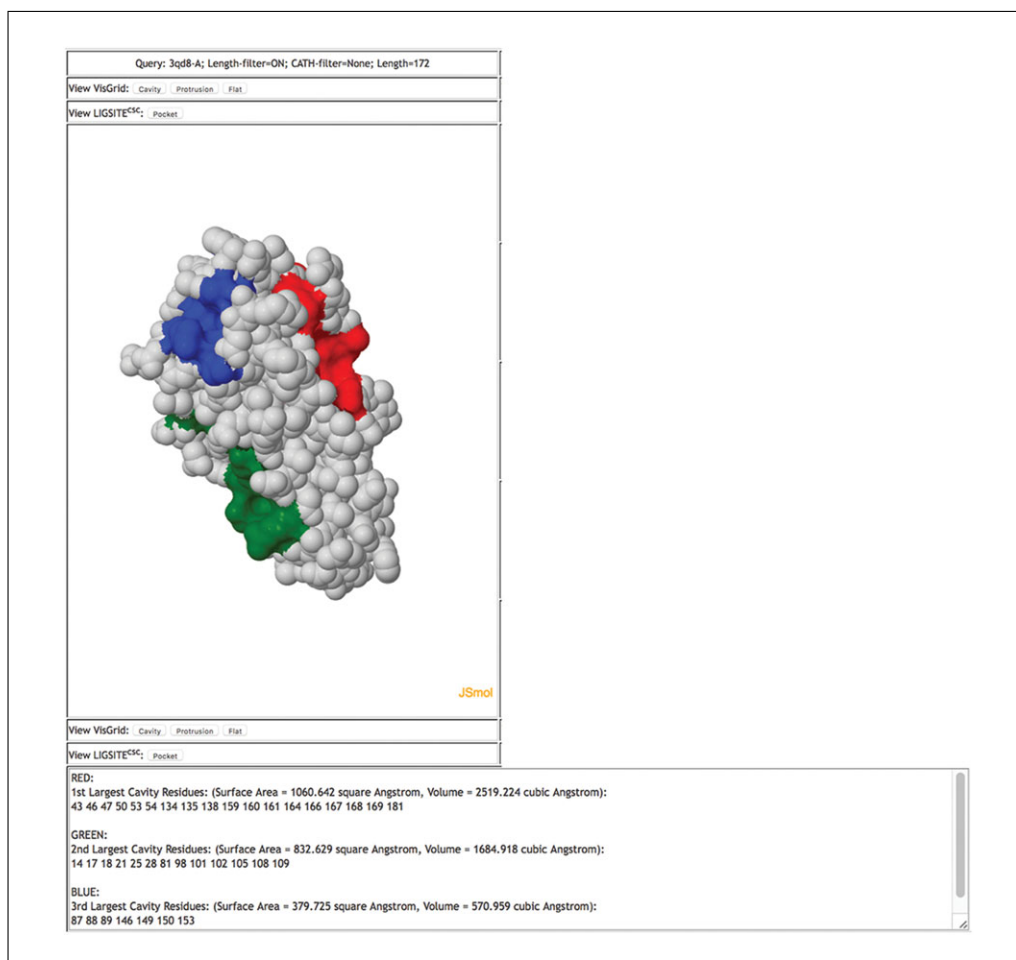


Figure 3.14.3 Geometric analysis of a query protein. The top part shows the query ID, filters used in the search, length of the query protein, and its CATH annotation if available. The query structure is visualized using the Jsmol applet. Users can click the Cavity, Protrusion, and Flat button to identify those regions using VisGrid. The first, second, and third largest cavities in this example are colored in red, green, and blue, respectively. The bottom panel lists residues in each cavity as well as its surface area and the volume.

residues will be colored using the same color scheme used for results from VisGrid. The surface area and volume of each pocket is also calculated with Qhull.

Structure of Retrieved Results

Retrieved structures are sorted according to the Euclidean distance between their 3DZD and the 3DZD of the query, which is shown as “EucD” in each panel (Fig. 3.14.4). In this example, the top 19 retrieved structures are ferritin homologs of BrfB with a Euclidean distance of 2.261 or less.

By moving the mouse over the image of a protein, it will rotate 360° along the *x* and *y* axes to give a through representation of the protein. Each retrieved protein is annotated by its structure ID, length, and CATH ID if available. The structure ID is linked to the corresponding entry in the PDB Web site.

Users can also run structure alignment between query and a retrieved protein by checking the “Rmsd” box below that structure. The alignment is performed using the Combinatorial Extension (CE) program (Shindyalov & Bourne, 1998). The RMSD value and the coverage (the number of aligned amino acids divided by the length of the query entry) will be displayed and a new Rmsd button will appear. By clicking the Rmsd button, the structure alignment is displayed using Jsmol applet on the left panel. To visualize



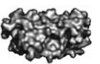
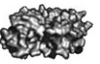



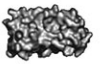
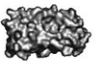



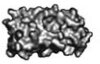
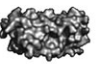



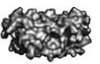
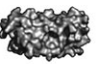




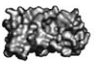

Results				
Top results in text format: 25				
Show				
 <u>3uno-C(169)</u> EucD: 1.683 CATH: N/A Rmsd: <input type="checkbox"/> 0.31Å (0.93)	 <u>3uno-A(175)</u> EucD: 1.699 CATH: N/A Rmsd: <input type="checkbox"/>	 <u>3uno-R(170)</u> EucD: 1.759 CATH: N/A Rmsd: <input type="checkbox"/>	 <u>3uno-N(172)</u> EucD: 1.783 CATH: N/A Rmsd: <input type="checkbox"/>	 <u>3uno-B(169)</u> EucD: 1.804 CATH: N/A Rmsd: <input type="checkbox"/>
 <u>3uno-T(168)</u> EucD: 1.919 CATH: N/A Rmsd: <input type="checkbox"/>	 <u>3uno-G(170)</u> EucD: 1.931 CATH: N/A Rmsd: <input type="checkbox"/>	 <u>3uno-L(168)</u> EucD: 1.970 CATH: N/A Rmsd: <input type="checkbox"/>	 <u>3uno-H(168)</u> EucD: 1.983 CATH: N/A Rmsd: <input type="checkbox"/>	 <u>3uno-W(168)</u> EucD: 2.024 CATH: N/A Rmsd: <input type="checkbox"/>
 <u>3uno-U(169)</u> EucD: 2.060 CATH: N/A Rmsd: <input type="checkbox"/>	 <u>3uno-F(168)</u> EucD: 2.079 CATH: N/A Rmsd: <input type="checkbox"/>	 <u>3uno-I(168)</u> EucD: 2.081 CATH: N/A Rmsd: <input type="checkbox"/>	 <u>3uno-Q(168)</u> EucD: 2.096 CATH: N/A Rmsd: <input type="checkbox"/>	 <u>3uno-O(168)</u> EucD: 2.111 CATH: N/A Rmsd: <input type="checkbox"/>
 <u>3uno-V(168)</u> EucD: 2.168 CATH: N/A Rmsd: <input type="checkbox"/>	 <u>3uno-K(168)</u> EucD: 2.176 CATH: N/A Rmsd: <input type="checkbox"/>	 <u>3uno-J(168)</u> EucD: 2.202 CATH: N/A Rmsd: <input type="checkbox"/>	 <u>3uno-X(167)</u> EucD: 2.261 CATH: N/A Rmsd: <input type="checkbox"/>	 <u>1cnt-2(130)</u> EucD: 2.285 CATH: 1.20.1250.10 Rmsd: <input type="checkbox"/>
 <u>3qd8-X(162)</u> EucD: 2.290 CATH: N/A Rmsd: <input type="checkbox"/>	 <u>3qd8-J(162)</u> EucD: 2.292 CATH: N/A Rmsd: <input type="checkbox"/>	 <u>3qd8-N(162)</u> EucD: 2.301 CATH: N/A Rmsd: <input type="checkbox"/>	 <u>3qd8-W(162)</u> EucD: 2.304 CATH: N/A Rmsd: <input type="checkbox"/>	 <u>3uno-M(168)</u> EucD: 2.308 CATH: N/A Rmsd: <input type="checkbox"/>
Top results in text format: 25				
Show				

Figure 3.14.4 Illustration of the top 25 retrieved structures in 3D-SURFER. Each hit is displayed with its structure ID, length, Euclidean distance to the query, and CATH classification if available. To calculate root mean squared deviation (RMSD) between the query and a specific hit, users can click on the checkbox following “Rmsd.” In this example, the RMSD between 3qd8-A and 3uno-C is 0.31 Å, and coverage is 93%. A list of the top 20, 50, 100, 250, 500, and 1000 retrieved structures can be displayed by specifying at the drop-down menu at top and clicking the Show button.

detailed alignment results, users can click on the RMSD result [“0.31Å (0.93)” in this example] and analyze the alignment file displayed in a new pop-up window.

Results for the top 25, 50, 100, 250, 500, and 1000 retrieved structures are available by choosing from the drop-down menu next to “Top results in text format.” After clicking Show, it will display structure ID, Euclidean distance, CATH ID, and length for proteins within the cutoff in a new window.

At the bottom of the results page, a line chart and exact numbers of the 3DZD (121 invariants) for the query protein (Fig. 3.14.5) are displayed.

Examples of Results Retrieved by 3D-SURFER

Here we show two examples of search results from 3D-SURFER. As 3DZD is capable of retrieving proteins with similar global surface, it can identify functionally related proteins with low sequence identity and insignificant structure similarity. 1a31-A is the structure of DNA topoisomerase I in human (Redinbo, Stewart, Kuhn, Champoux, & Hol, 1998).

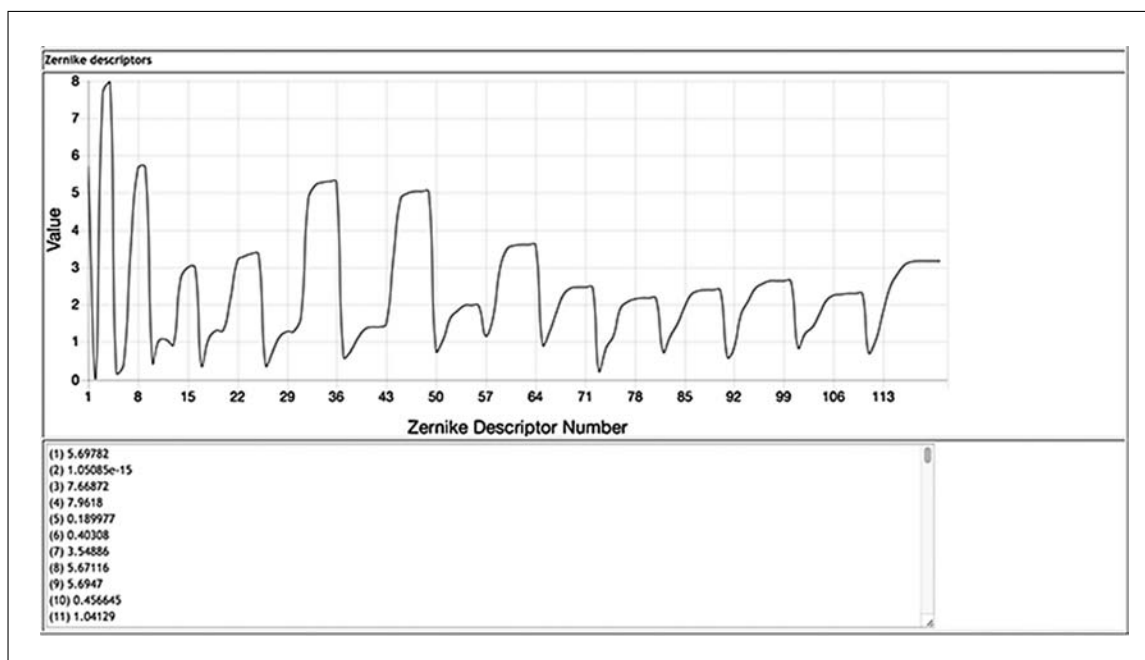


Figure 3.14.5 Graphic and text representation of 3DZD for a query protein. In this example, it displays 3DZD for 3qd8-A.

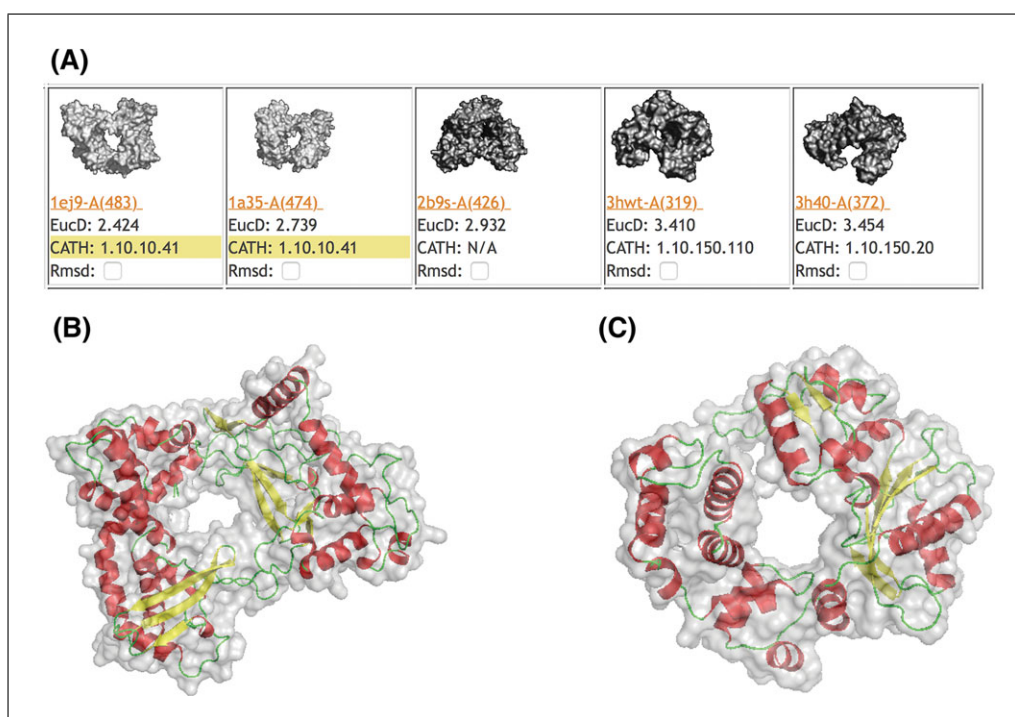


Figure 3.14.6 An example of a protein pair with similar global surface but different folds. **(A)** Part of search results for 1a31-A on 3D-SURFER. The top 5 hits are shown. **(B)** 1a31-A, DNA topoisomerase I from human. **(C)** 3hwt-A, DNA polymerase lambda from human.

Using 1a31-A as query in 3D-SURFER, the top three most similar structures (PDB ID: 1ej9-A, 1a35-A, 2b9s-A) are DNA topoisomerase I in human and *Leishmania donovani* with Euclidean distances less than 3 (Fig. 3.14.6A). The default search setting was used in this example. The structure retrieved at rank 4 is DNA polymerase lambda (3hwt-A) in human. Similar to DNA topoisomerase I, DNA polymerase lambda shares a characteristic central pore that binds to DNA double strands (Fig. 3.14.6B and 3.14.6C). The sequence identity between 1a31-A and 3hwt-A is low (21.6%), and the RMSD between the two

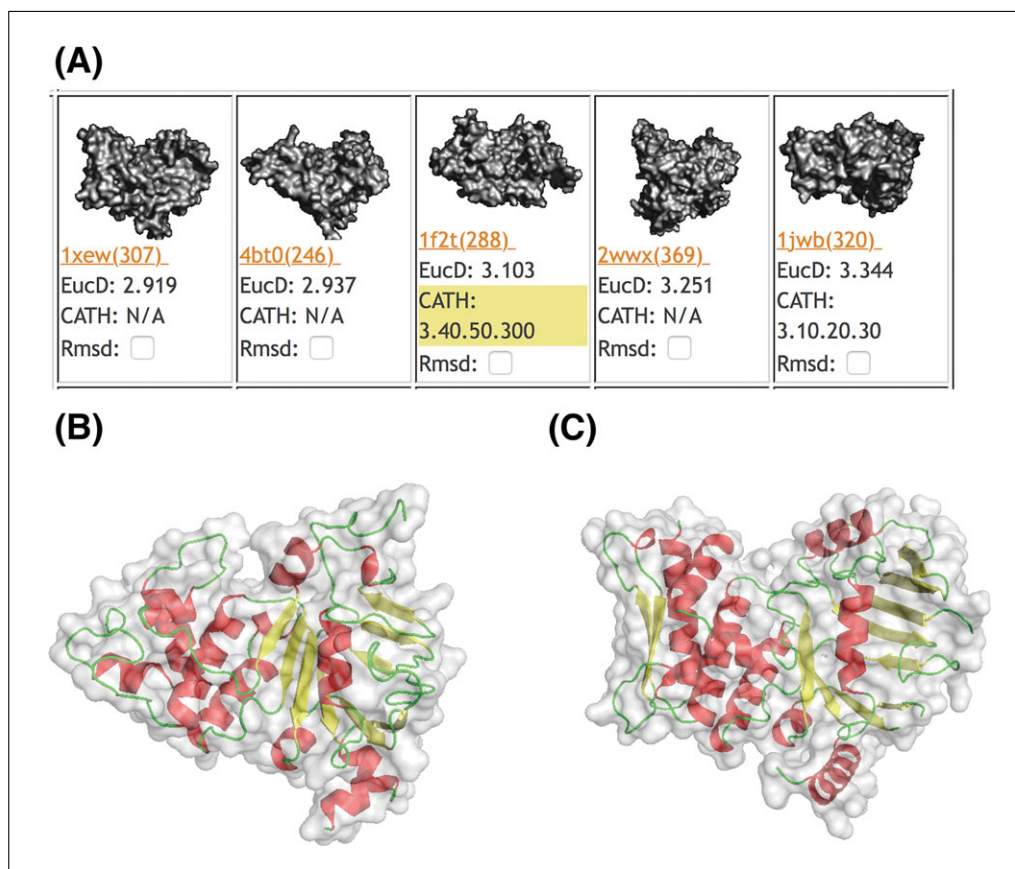


Figure 3.14.7 An example of search results for ATPase domain in TAP1 (PDB ID: 2ixf-A) against the complex database. **(A)** Part of search results for 2ixf-A by 3D-SURFER. The top 5 hits are shown. **(B)** Structure of query protein 2ixf-A, ATPase domain in TAP1 from *Rattus norvegicus*. **(C)**, Structure of the top hit, 1xew, SMCcd-SMCcd homodimer from *Pyrococcus furiosus*.

structures is 9.6 Å. However, 3DZD is able to identify their overall surface similarity with a 3.41 Euclidean distance.

The second example is a search for a single chain protein query against the protein complex database. All other parameters were set to default in this search. 2ixf-A is the ATPase domain of TAP1, a subunit of ATP-binding cassette (ABC) transporter TAP in *Rattus norvegicus* (Procko, Ferrin-O'Connell, Ng, & Gaudet, 2006; Fig. 3.14.7). The top three retrieved structures (Fig. 3.14.7A) are all dimeric complexes with the ABC-ATPase fold, which have an Euclidean distance less than 3.2 to the query. The first retrieved complex is the SMCcd-SMCcd homodimer from *Pyrococcus furiosus* (1xew; Lammens, Schele, & Hopfner, 2004; Fig. 3.14.7C). An SMC protein has a catalytic ATP binding cassette (ABC) domain with the ATPase activity. It has a similar global surface to 2ixf-A as reflected in its Euclidean distance, but different secondary structure arrangements, which give an RMSD of 6.24 Å.

EM-SURFER Results

A search result of EM-SURFER is displayed in a similar layout as in 3D-SURFER. The results page displays the top 20 structures that share a similar global isosurface shape to the input map. Similar to 3D-SURFER, global surface similarity between two maps is quantified by the Euclidean distance of their 3DZDs. The smaller the distance, the more similar the two EM maps are. Empirically, two biomolecules in EM maps are biologically related if the distance is smaller than 8.0. Below, we explain and discuss search results of the EM-SURFER server using EMD-1180 as a query, which is a GroEL-ATP-GroES complex (Ranson et al., 2006), as shown in Figure 3.14.8.

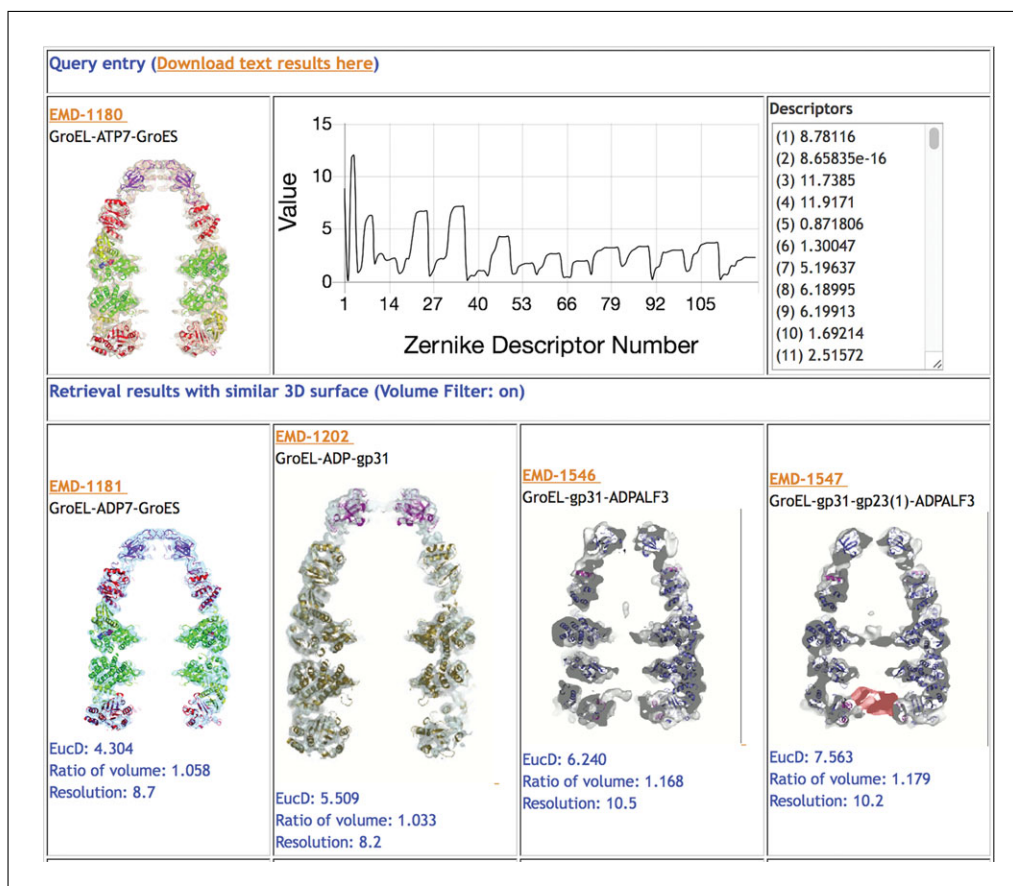


Figure 3.14.8 An example of the EM-SURFER results page for query EMD-1180. The top left panel displays an image of the query structure, or the filename if a user's EM map is uploaded. The graph and the text depiction of the 3DZD for the query are shown next to the map image. The search results panel shows top 20 hits with their EMD ID, a short description, structure image, and detailed values of their Euclidean distance, volume ratio, and resolution. Here we only show part of the results page (top 4 hits) for EMD-1180.

The top panel in the results page displays the EMD ID of the query, its description, and a figure showing its overall structure, which is taken from EMD. Graphic and text forms of the 3DZD of the query protein are placed next to the structure image. To download text results for the top 50 structures, users can click the "Download text results here" button located in the first row. In the text results, retrieved structures are ranked by Euclidean distances of their 3DZD to the query. Resolution information is also provided, if available, in the map information in EMD.

The search results panel shows the top 20 structures with the most similar global iso-surface shape to the query (Fig. 3.14.8). In this example, the top 13 retrieved EM maps are all GroELs, which indicate that the search is successful in identifying biologically relevant maps to the query from EMD. Retrieved structures are ranked by the Euclidean distance of their 3DZDs to that of the query (shown after "EucD:"). Smaller Euclidean distance indicates that the surface of two structures is more similar. Each hit is displayed with its EMD ID, a short description, the ratio of volume to query, and its resolution. The EMD ID of the structure is linked to its corresponding entry in the EMD Web site to allow users to obtain more detailed structure information. By clicking on the image of a retrieved structure, users can start a new search from the clicked entry. Default filter settings are used in this new search.

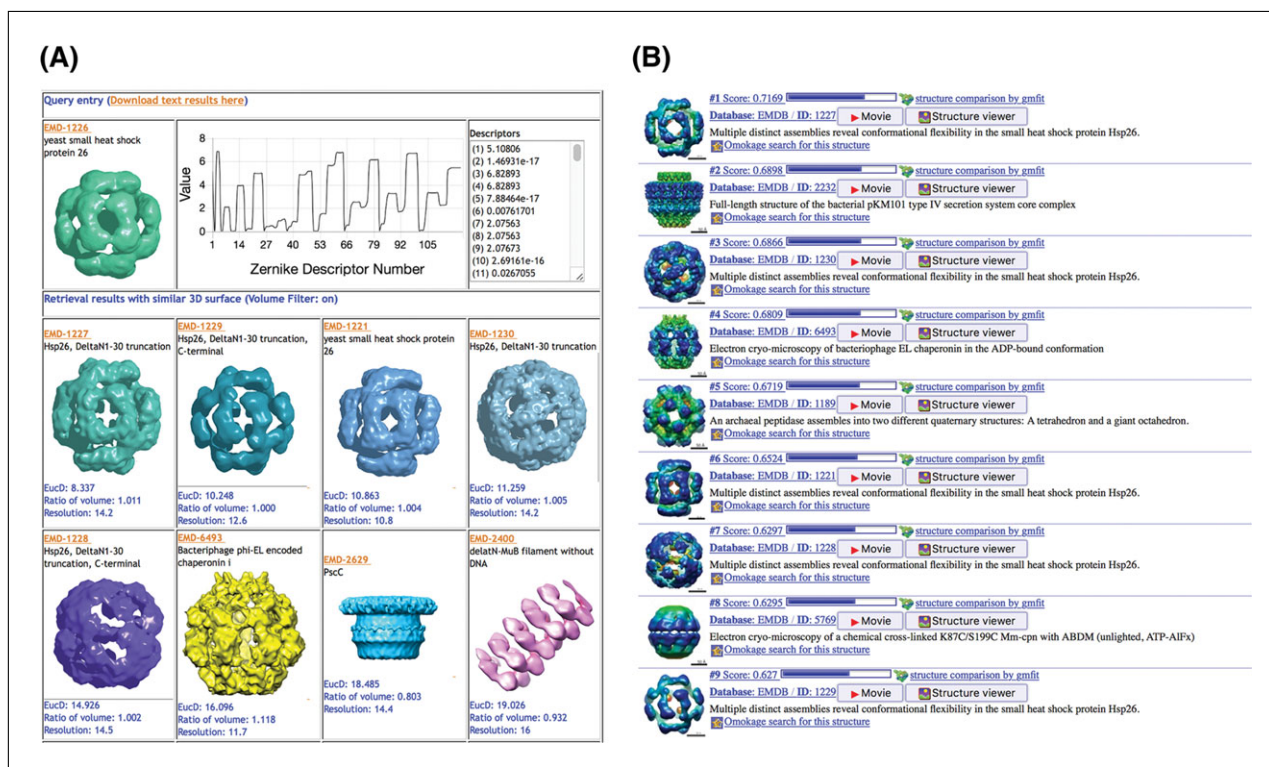


Figure 3.14.9 Search results for a yeast heat shock protein Hsp26, EMD-1226. **(A)** Search results in EM-SURFER. The top 8 hits are shown. EMD-1227, 1229, 1221, 1230, and 1228 are EM maps for Hsp26. **(B)** Search results in the Omokage server. EMD-1227, 1230, 1221, 1228, and 1229 are EM maps for Hsp26.

An Example of Retrieval by EM-SURFER

Here we show another search result by EM-SURFER, using EMD-1226 as the query (Fig. 3.14.9A). For this search, the author-recommended contour level was used and the volume filter was on. The query is yeast heat shock protein Hsp26 in compact form (White et al., 2006). Six different three-dimensional structures of wild-type Hsp26 and modified forms were solved in the original paper (EMD-1221, 1226 to 1230). All of them are retrieved at the top in the search results. There are two distinct forms of Hsp26, one in a compact and another in an expanded form. The internal organization of those two forms is different, but their external diameter is only 4% different. As both forms share similar global surface shape, using EMD-1226 as the query, all other Hsp26 proteins are ranked top 1 to top 5 in the EM-SURFER results. For comparison, in Figure 3.14.9B we show the search results from the Omokage server, another tool for searching PDB and EMD (Suzuki, Kawabata, & Nakamura, 2016). It uses a combination of incremental distance rank (iDR) profile and the principal component analysis (PCA) profile to characterize shape similarity. Although five other Hsp26 proteins are retrieved among top nine hits on the Omokage server, they are separated by some other proteins, like bacteriophage EL chaperonin. Thus, in this case, EM-SURFER had a better performance than Omokage.

COMMENTARY

Background Information

3D Zernike descriptors (3DZD)

3DZD are based on a mathematical series expansion of a given 3D function. They have been applied in various comparisons of biomolecular data (Kihara et al., 2011), including protein-protein docking (Esquivel-Rodriguez, Yang, & Kihara, 2012;

Li & Kihara, 2012; Venkatraman, Yang, Sael, & Kihara, 2009), ligand binding pocket comparison (Chikhi, Sael, & Kihara, 2010; Sael & Kihara, 2012), and ligand molecule search (Venkatraman, Chakravarthy, & Kihara, 2009). It is rotation invariant, meaning that prior structure alignment is not required for calculation of 3DZD.

In 3D-SURFER, the calculation of 3DZD starts by construction of surface triangulation using the MSROLL (Connolly, 1993) and MSMS programs (Sanner, Olson, & Spehner, 1996). The constructed triangle mesh is then mapped to a 3D grid. Voxels that overlap with protein surface are assigned the value 1, and 0 otherwise. This discrete representation is used as the input function $f(x)$ for 3DZD calculation. In EM-SURFER, depending on the contour level specified, voxels with electron density equal to or larger than the contour level are marked with 1, and 0 otherwise. This value-mapped 3D grid is taken as the input function $f(x)$ for 3DZD calculation.

The 3D function $f(x)$ is expanded into a series in terms of Zernike-Canterakis basis (Novotni & Klein, 2003) defined as follows:

$$\Omega_{nl}^m = \frac{3}{4\pi} \int_{|x| \leq 1} f(x) \bar{Z}_{nl}^m(x) dx,$$

where $Z_{nl}^m(r, \theta, \Phi) = R_{nl}(r) Y_l^m(\theta, \Phi)$

Equation 3.14.1

The ranges of parameters m and l depend on order n : $-l < m < l$, $0 \leq l \leq n$, and $(n-l)$ is even. Previous study has shown that order $n = 20$ offers a sufficiently accurate representation (Novotni & Klein, 2003). Therefore, we set $n = 20$, which generates 121 invariants. Here $Y_l^m(\theta, \Phi)$ are spherical harmonics (Dym & McKean, 1972) and $R_{nl}(r)$ are the radial functions defined by Canterakis. The rotation-invariant 3D Zernike descriptors are calculated as norms of Ω_{nl}^m :

$$F_{nl} = \sqrt{\sum_{m=-l}^{m=l} (\Omega_{nl}^m)^2}$$

Equation 3.14.2

The similarity between two sets of 3DZDs is quantified by their Euclidean distance d_E :

$$d_E = \sqrt{\sum_{i=1}^{121} (X_i - Y_i)^2}$$

Equation 3.14.3

where X_i and Y_i represent the i th invariant for each protein.

Critical Parameters

As proteins are scaled into a unit sphere for calculation of 3DZD, the size information of a protein is not reflected in its 3DZD (Novotni &

Klein, 2003). However, our previous work has shown that it is uncommon for proteins with very different sizes to have the same global surface shape (Sael et al., 2008). Therefore, in 3D-SURFER, turning off the length filter may not change search results significantly.

In EM-SURFER, when a user uploads an EM map, the key parameter affecting search accuracy is the contour level used in construction of the isosurface shape. As the search is based on global surface shape, different contour levels will change the global surface significantly and affect accuracy of the search results. If a retrieved result does not include maps that are known to be similar to the query map, try searching with a different contour level option. Note that EM-SURFER identifies maps that share similar global isosurface shape. Thus, maps of the same protein may not be retrieved if they are in different biological states that have significantly different protein conformations or different bound proteins/molecules.

Troubleshooting

When a query ID is not recognized in the structure ID box in 3D-SURFER, it is possible that the structure entered is obsolete in PDB or the structure sequence is too short. PDB entries can be made obsolete following an author's request to PDB when better experimental data has been collected or a better interpretation of existing data has been produced. Obsolete entries reported in PDB are removed from the 3D-SURFER database, making those structure IDs unrecognizable in the search. In this case, users can go to the PDB Web page, identify a superseding entry for that obsolete structure, and input its successor into the search. Another possibility is that the input structure contains less than 10 residues. Those short structures are also removed in 3D-SURFER, as their structure information is limited.

When an error occurs in the uploading process in EM-SURFER, it may come from a network connectivity issue, incorrect format in map file, etc. EM-SURFER has several checks and repair mechanisms to process uploaded map files. However, if there is still problem with map uploading, users are welcome to contact the authors via e-mail (dkihara@purdue.edu).

Acknowledgements

We acknowledge Kevin Shim for proof-reading the manuscript. This work was supported by National Institutes of Health (R01GM123055) and National Science

Literature Cited

- Andreeva, A., Howorth, D., Chothia, J. M., Brenner, S. E., Hubbard, T. J., Chothia, C., & Murzin, A. G. (2008). Data growth and its impact on the SCOP database: New developments. *Nucleic Acids Research*, 36 (Database issue), D419–425. doi: 10.1093/nar/gkm993.
- Andreeva, A., Howorth, D., Chothia, C., Kulesha, E., & Murzin, A. G. 2015. Investigating protein structure and evolution with SCOP2. *Current Protocols in Bioinformatics*, 49, 1.26.1–1.26.21. doi: 10.1002/0471250953.bi0126s49.
- Barber, C. B., Dobkind, D. P., & Huhdanpaa, H. (1996). The Quickhull algorithm for convex hulls. *ACM T Math Software*, 22(4), 469–483. doi: 10.1145/235815.235821.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., ... Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1), 235–242. doi: 10.1093/nar/28.1.235.
- Chikhi, R., Sael, L., & Kihara, D. (2010). Real-time ligand binding pocket database search using local surface descriptors. *Proteins*, 78(9), 2007–2028. doi: 10.1002/prot.22715.
- Connolly, M. L. (1993). The molecular surface package. *Journal of Molecular Graphics*, 11(2), 139–141. doi: 10.1016/0263-7855(93)87010-3.
- Di Costanzo, L., Ghosh, S., Zardecki, C., & Burley, S. K. (2016). Using the tools and resources of the RCSB protein data bank. *Current Protocols in Bioinformatics*, 55, 1.9.1–1.9.35. doi: 10.1002/cpbi.13.
- Dym, H., & McKean, H. P. (1972). *Fourier series and integrals*. New York: Academic Press.
- Esquivel-Rodriguez, J., Xiong, Y., Han, X., Guang, S., Christoffer, C., & Kihara, D. (2015). Navigating 3D electron microscopy maps with EM-SURFER. *BMC Bioinformatics*, 16, 181. doi: 10.1186/s12859-015-0580-6.
- Esquivel-Rodriguez, J., Yang, Y. D., & Kihara, D. (2012). Multi-LZerD: Multiple protein docking for asymmetric complexes. *Proteins*, 80(7), 1818–1833. doi: 10.1002/prot.24079.
- Huang, B., & Schroeder, M. (2006). LIGSITEcsc: Predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Structural Biology*, 6, 19. doi: 10.1186/1472-6807-6-19.
- Khare, G., Gupta, V., Nangpal, P., Gupta, R. K., Sauter, N. K., & Tyagi, A. K. (2011). Ferritin structure from *Mycobacterium tuberculosis*: Comparative study with homologues identifies extended C-terminus involved in ferroxidase activity. *PLoS One*, 6(4), e18570. doi: 10.1371/journal.pone.0018570.
- Kihara, D., Sael, L., Chikhi, R., & Esquivel-Rodriguez, J. (2011). Molecular surface representation using 3D Zernike descriptors for protein shape comparison and docking. *Current Protein & Peptide Science*, 12(6), 520–530. doi: 10.2174/138920311796957612.
- La, D., Esquivel-Rodriguez, J., Venkatraman, V., Li, B., Sael, L., Ueng, S., ... Kihara, D. (2009). 3D-SURFER: Software for high-throughput protein surface comparison and analysis. *Bioinformatics*, 25(21), 2843–2844. doi: 10.1093/bioinformatics/btp542.
- Lammens, A., Schele, A., & Hopfner, K. P. (2004). Structural biochemistry of ATP-driven dimerization and DNA-stimulated activation of SMC ATPases. *Current Biology*, 14(19), 1778–1782. doi: 10.1016/j.cub.2004.09.044.
- Lawson, C. L., Patwardhan, A., Baker, M. L., Hryc, C., Garcia, E. S., Hudson, B. P., ... Chiu, W. (2016). EMDDataBank unified data resource for 3DEM. *Nucleic Acids Research*, 44(D1), D396–403. doi: 10.1093/nar/gkv1126.
- Li, B., & Kihara, D. (2012). Protein docking prediction using predicted protein-protein interface. *BMC Bioinformatics*, 13(7). doi: 10.1186/1471-2105-13-7.
- Li, B., Turuvekere, S., Agrawal, M., La, D., Ramani, K., & Kihara, D. (2008). Characterization of local geometry of protein surfaces with the visibility criterion. *Proteins*, 71(2), 670–683. doi: 10.1002/prot.21732.
- Novotni, M., & Klein, R. (2003). *3D Zernike descriptors for content based shape retrieval*. SM '03 Proceedings of the eighth ACM symposium on solid modeling and applications, Seattle, Washington—June 16–20, 2003 (pp. 216–225). Washington DC: American Chemical Society.
- Procko, E., Ferrin-O'Connell, I., Ng, S. L., & Gaudet, R. (2006). Distinct structural and functional properties of the ATPase sites in an asymmetric ABC transporter. *Molecules and Cells*, 24(1), 51–62. doi: 10.1016/j.molcel.2006.07.034.
- Ranson, N. A., Clare, D. K., Farr, G. W., Houldershaw, D., Horwich, A. L., & Saibil, H. R. (2006). Allosteric signaling of ATP hydrolysis in GroEL-GroES complexes. *Nature Structural & Molecular Biology*, 13(2), 147–152. doi: 10.1038/nsmb1046.
- Redinbo, M. R., Stewart, L., Kuhn, P., Champoux, J. J., & Hol, W. G. (1998). Crystal structures of human topoisomerase I in covalent and noncovalent complexes with DNA. *Science*, 279(5356), 1504–1513. doi: 10.1126/science.279.5356.1504.
- Sael, L., & Kihara, D. (2010). Improved protein surface comparison and application to low-resolution protein structure data. *BMC Bioinformatics*, 11 (Suppl 11), S2. doi: 10.1186/1471-2105-11-S11-S2.
- Sael, L., & Kihara, D. (2012). Detecting local ligand-binding site similarity in nonhomologous proteins by surface patch comparison. *Proteins*, 80(4), 1177–1195. doi: 10.1002/prot.24018.
- Sael, L., Li, B., La, D., Fang, Y., Ramani, K., Rustamov, R., & Kihara, D. (2008). Fast protein tertiary structure retrieval based on global

- surface shape similarity. *Proteins*, 72(4), 1259–1273. doi: 10.1002/prot.22030.
- Sanner, M. F., Olson, A. J., & Spehner, J. C. (1996). Reduced surface: An efficient way to compute molecular surfaces. *Biopolymers*, 38(3), 305–320. doi: 10.1002/(SICI)1097-0282(199603)38:3<305::AID-BIP4>3.0.CO;2-Y.
- Shindyalov, I. N., & Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering*, 11(9), 739–747. doi: 10.1093/protein/11.9.739.
- Sillitoe, I., Lewis, T. E., Cuff, A., Das, S., Ashford, P., Dawson, N. L., . . . Orengo, C. A. (2015). CATH: Comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Research*, 43 (Database issue), D376–381. doi: 10.1093/nar/gku947.
- Suzuki, H., Kawabata, T., & Nakamura, H. (2016). Omokage search: Shape similarity search service for biomolecular structures in both the PDB and EMDb. *Bioinformatics*, 32(4), 619–620. doi: 10.1093/bioinformatics/btv614.
- Venkatraman, V., Chakravarthy, P. R., & Kihara, D. (2009). Application of 3D Zernike descriptors to shape-based ligand similarity searching. *Journal of Cheminformatics*, 1, 19. doi: 10.1186/1758-2946-1-19.
- Venkatraman, V., Yang, Y. D., Sael, L., & Kihara, D. (2009). Protein-protein docking using region-based 3D Zernike descriptors. *BMC Bioinformatics*, 10, 407. doi: 10.1186/1471-2105-10-407.
- White, H. E., Orlova, E. V., Chen, S., Wang, L., Ignatiou, A., Gowen, B., . . . Saibil, H. R. (2006). Multiple distinct assemblies reveal conformational flexibility in the small heat shock protein Hsp26. *Structure*, 14(7), 1197–1204. doi: 10.1016/j.str.2006.05.021.
- Wilson, C. A., Kreychman, J., & Gerstein, M. (2000). Assessing annotation transfer for genomics: Quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *Journal of Molecular Biology*, 297(1), 233–249. doi: 10.1006/jmbi.2000.3550.
- Xiong, Y., Esquivel-Rodriguez, J., Sael, L., & Kihara, D. (2014). 3D-SURFER 2.0: Web platform for real-time search and characterization of protein surfaces. *Methods in Molecular Biology*, 1137, 105–117. doi: 10.1007/978-1-4939-0366-5_8.