

3D-SURFER 2.0: Web Platform for Real-Time Search and Characterization of Protein Surfaces

Yi Xiong, Juan Esquivel-Rodriguez, Lee Sael, and Daisuke Kihara

Abstract

The increasing number of uncharacterized protein structures necessitates the development of computational approaches for function annotation using the protein tertiary structures. Protein structure database search is the basis of any structure-based functional elucidation of proteins. 3D-SURFER is a web platform for real-time protein surface comparison of a given protein structure against the entire PDB using 3D Zernike descriptors. It can smoothly navigate the protein structure space in real-time from one query structure to another. A major new feature of Release 2.0 is the ability to compare the protein surface of a single chain, a single domain, or a single complex against databases of protein chains, domains, complexes, or a combination of all three in the latest PDB. Additionally, two types of protein structures can now be compared: all-atom-surface and backbone-atom-surface. The server can also accept a batch job for a large number of database searches. Pockets in protein surfaces can be identified by VisGrid and LIGSITE^{cs}. The server is available at <http://kiharalab.org/3d-surfer/>.

Key words Structure-based function prediction, Protein surface, Surface shape comparison, Structure similarity, 3D Zernike descriptor, database search

1 Introduction

With the progress of structural genomics initiatives, an increasing proportion of solved protein structures are not functionally annotated. It remains a challenging task to understand the relationship between protein structure and function and to extrapolate a working mechanism of cellular machinery. To this end, a large number of computational approaches have been developed for function annotation based on protein sequences and their tertiary structures [1, 2]. Structure-based approaches are advantageous over sequence-based methods in the sense that distant relationships of proteins can be better identified by structure comparison, which can also often identify functional residues that are not localized in their amino acid sequences. Analogous to the important role of

BLAST [3] in sequence-based function annotation methods, protein global structure searches against the Protein Data Bank (PDB) [4] are the basis of any structure-based functional elucidation of proteins.

An intuitive approach for comparing two protein structures is to align the atoms or residues of the proteins. However, the structural alignment procedures are time consuming, making it unfeasible for searching against the entire protein structure database in real-time. To speed up the database search, numerous alignment-free methods have been developed. For example, 3D-BLAST encodes the structure as a 1D sequence of alphabets [5]. The LightField Descriptor is constructed by 2D projections (a combination of 2D Zernike descriptors and Fourier coefficients) rendered from uniformly distributed points around a sphere that surrounds the protein surface [6]. Unlike the methods that are based on 1D or 2D representations, 3D moment-based shape representations truly capture the 3D geometrical shape of proteins. Among them, the 3D Zernike descriptor (3DZD) has been shown to be suitable for the efficient comparison of protein surfaces [7]. 3DZD is a series expansion of a 3D mathematical function and thus can represent a 3D object compactly as a vector of coefficients of the series function. Moreover, 3DZD is rotation invariant, which means that structure alignment is not needed prior to comparison. These two characteristics allow for fast, real-time searches of structure databases. 3DZD has been successfully applied for fast comparisons of various biological structure data [8] including ligand binding pockets [9, 10], low-resolution electron microscopy data [11], ligand molecules [12], and protein-protein docking [13, 14].

Here, we present 3D-SURFER 2.0, an upgraded web-based platform for high-throughput protein surface comparison, analysis, and visualization [15]. The server compares the protein surface of a single chain, a single domain, or a single complex against databases of protein chains, domains, complexes, or a combination of all three in the latest PDB. By using the 3DZD representation of the structures, the search process will be completed in a couple of seconds. A query structure can be selected by its PDB code or uploaded from a local computer of a user. In addition, local geometrical characteristics of a query protein such as pocket regions can be identified by VisGrid [16] or LIGSITE^{sc} [17]. Retrieved structures for a query are visualized with animations of rotating proteins. Clicking a protein icon in a search result will invoke another search from the clicked structure to surf into the protein structure space. Structures are associated with CATH codes [18] and conventional root mean square deviation (RMSD); main-chain structure alignment can be computed with the combinatorial extension (CE) algorithm [19].

2 Materials

2.1 Input Data

The input data of 3D-SURFER is a protein structure, which will be compared against a user-specified dataset of the entire PDB database. The input structure is provided by entering its identification (ID) code in the search window or by uploading a PDB format file to the server. The ID code of an input protein structure is named based on the PDB ID of the protein. If an entire structure (e.g., a protein complex) in a PDB file is chosen for input, the ID is the same as the PDB code (e.g., 7tim). A chain in a PDB file can be specified by adding a hyphen and the chain ID following the PDB code (e.g., 7tim-A). A domain of a chain can be specified by further adding a domain ID that is defined by CATH (e.g., 7tim-A-01). In each case, a search against the entire structure database is executed on-the-fly.

Two options are provided for protein surface representation: the surface is generated from all surface atoms or is constructed using only backbone atoms. This is because our previous work [11] showed that depending on the query structure, one of the surface representations will agree better in structure retrieval to the conventional fold classification (CE [19] and SCOP [20]). For a target database to be searched against, four types of datasets are prepared: protein single chains, domains, complexes, or a combination of the three. Additionally, the user can specify two types of filters: a CATH filter that avoids displaying multiple structures with the same CATH level and a length filter which retrieves only proteins whose lengths are similar to the query structure.

2.2 Programs Used in the Server

Individual programs of the server are seamlessly integrated behind the scenes and thus will not be visible to users of 3D-SURFER. Here, we briefly describe the programs and computational steps implemented within the server. The steps to calculate 3DZD are summarized as follows. First, the protein molecular surface is triangulated by MSROLL [21] (*see Note 1*). The extracted mesh is then discretized to generate a cubic grid (a binary voxelization, *see Note 2*). Finally, the 3DZD is computed [22] by taking the cubic grid as input (a vector of 121 invariants). An input protein structure represented as a vector can be compared with other structures by computing the Euclidean distance of the vectors (*see Note 3*). Moreover, the VisGrid and LIGSITE^{cs} programs are integrated for characterizing the geometry of local surface regions in a query structure. In addition, the conventional RMSD of C α atoms can be computed between a query structure and retrieved structures specified by clicking the “RMSD” button. The CE program is invoked for the RMSD computation (*see Note 4*). The query structure is visualized with the Jmol applet (<http://www.jmol.org/>) (*see Note 5*).

2.3 Web Sites

3D-SURFER version 2.0 is available at <http://kiharalab.org/3d-surfer/>. The previous version is still accessible at <http://kiharalab.org/3d-surfer/v1.0> or from a tab in the menu bar. The Web site has a navigation menu bar that includes tabs for a regular job submission, batch job submission, tutorial, reference, contact information, and 3D-SURFER version 1.0.

3 Methods

3.1 Submit a Query Entry

Submitting a query entry on the 3D-SURFER Web site consists of four steps. The details are explained below (Fig. 1).

The steps to submit an entry

Step 1: Input a structure ID or upload a structure file

Structure ID: <input type="text"/> e.g. Chain ID: 7tim-A Complex ID: 2wiw Domain ID: 1h41-B-02	Or	Upload a structure file: <input type="text"/> <input type="button" value="Browse..."/> (Optional) Please specify your domain range in your uploaded file: <input type="text"/>
---	----	---

Step 2: Select a surface atom representation

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

Step 3: Choose a database to compare

Template Database :	Chain <input type="button" value="v"/> Chain Complex Domain All of above
---------------------	--

Step 4: Specify CATH and Length filters

CATH filter:	CATH <input type="button" value="v"/>	<input type="radio"/> OFF
Length filter:	None CATH CAT CA C	

Fig. 1 The steps to submit a structure in 3D-SURFER

1. *Input a structure ID or upload a structure file:* In this step, users can either type a structure ID or upload a structure file. As shown in the example, the IDs should consist of four, six, or nine characters. When users type the first character, the text field will pop up a dropdown menu showing a list of IDs that start with the specified character. When the file upload option is used, the input should be consistent with the PDB format. At a minimum, the file should contain ATOM lines that contain atomic coordinates of proteins.
2. *Select a surface atom representation:* Two different surface representation options are provided: the all-atom-surface representation and the backbone-atom-surface representation, which includes C α , C, and N atoms in the main chain. The choice should be made according to the purpose of the structure database search. If proteins with similar main-chain orientations are sought, the backbone-atom-surface representation works best in many cases (except for proteins with floppy tails and unpacked structures, *see* our previous work [11]).
3. *Choose a database to be searched:* Users can compare the surface of a single protein chain to a dataset of single chains, domains, complexes, or a combination of the three in the latest PDB.
4. *Specify CATH and length filters:* It is common that retrieved results contain many homologous proteins or essentially the same proteins in different conditions. Users can filter out such similar structures using the CATH filter. If users specify CATH filters ("CATH", "CAT", "CA", "C", or no filter) in the dropdown menu, the returned search results will contain only one structure from the same CATH classification of the specified level. For example, if the CATH filter level "CAT" is specified, no proteins in the retrieved list share the first three digits. When the length filter is turned on, the search result will only contain proteins with similar length to the query, i.e., those which have a length ratio between 0.57 and 1.75 to the query protein. Note that the size information of proteins is lost in 3DZD. However, our work [7] shows that it is uncommon for proteins with very different sizes to have the same surface shape and thus turning on the length filter may not drastically change the search results.

3.2 Result Page

In this section, we will show how the surface comparison results are presented. The similarity of two proteins are quantified by the Euclidean distance (the square root of the sum of the squares of the differences between corresponding values) between the 3DZD vectors (121 scalar values) of the proteins that represent protein surface shape. In the 3D-SURFER results panel, the Euclidean distance is shown at the label "EucD:". If two structures have an Euclidean distance less than a threshold of 10, they can be considered significantly similar.

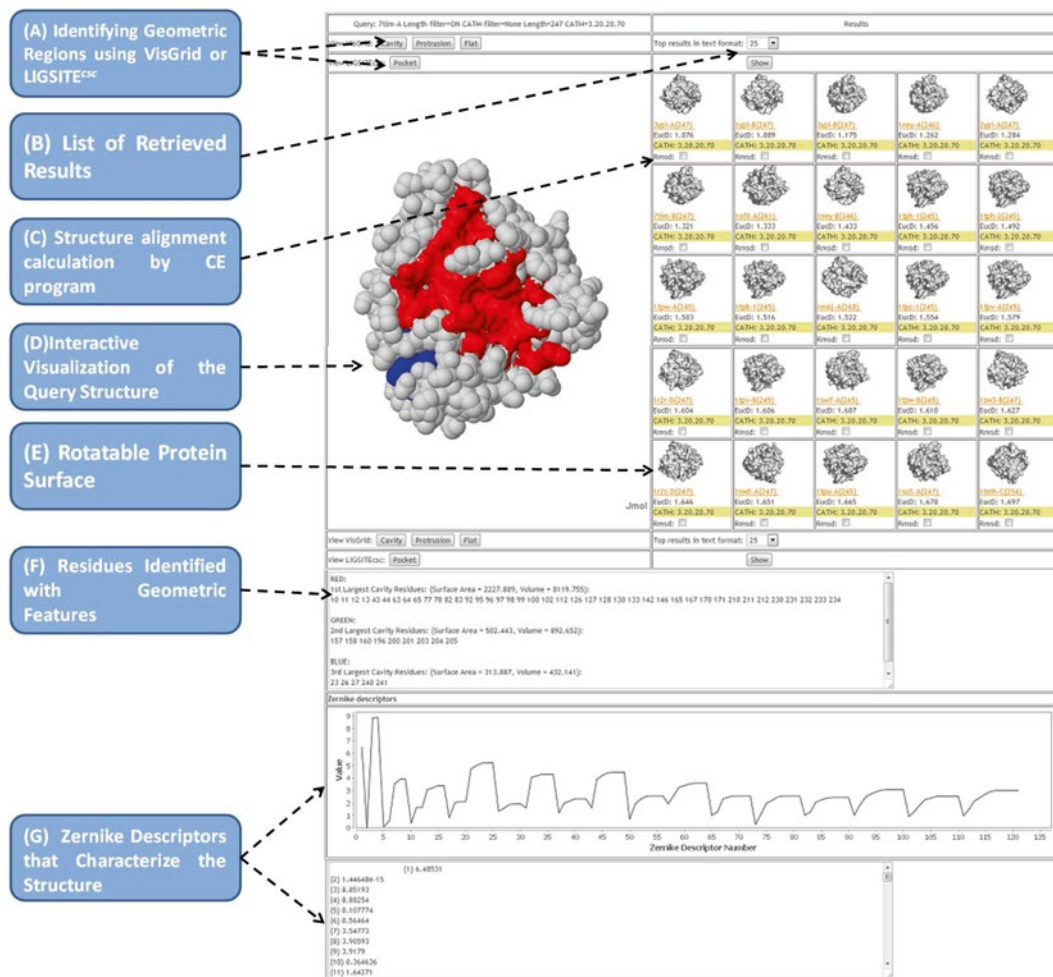


Fig. 2 The search result page of 3D-SURFER

1. *Viewing the query protein and its geometric features*: The query structure is visualized by the Jmol applet at the left side of the panel (Fig. 2d). By clicking the right button of the mouse, the representation of the protein can be changed using the functionality of Jmol. Geometrical features of the query structures, namely cavity, protrusion, and flat regions can be identified using VisGrid by clicking buttons located on the upper-left position of the panel (Fig. 2a). The “pocket” button in the third row invokes LIGSITE^{ESC} to identify pockets on the protein surface. Identified pockets, protrusions, and flat regions are indicated with colors on the protein surface, where the rank is based on the size (*see Note 6*) (Fig. 2a, d, f). Users can

also look at the volumes and surface areas (*see Note 7*) of the convex hull (*see Note 8*) formed from the atom coordinates of the residues identified by VisGrid or LIGSITE^{cs}.

2. *Protein structure retrieval results and further navigation of the protein surface universe*: Retrieved structures from the database are shown on the right side of the panel. They are ranked by the Euclidean distance of the 3DZD. By default, the top 25 structures are displayed but the number of structures to show can be changed to the top 50, 100, 250, 500, or 1,000 (Fig. 2b). Protein surfaces can be rotated by moving the mouse over animation images of proteins. The images will spin 360° along both the *X* and *Y* axes to present a complete view of the protein surface (Fig. 2e). Each retrieved structure is associated with the PDB ID and the CATH code, if assigned (some PDB entries are not indexed in the CATH database). The PDB ID is linked to the entry in the PDB Web site.

Clicking the image of a retrieved structure will invoke a new search from that structure against the database specified for the initial search. Welcome to the protein surface universe—users can enjoy surfing in the ocean of protein surfaces by taking advantage of real-time searches!

3. *Structure alignment calculations and visualization*: Conventional main-chain-based structure alignments can be performed using the CE program (*see Note 9*) to obtain RMSD values. When checking the "Rmsd:" box (Fig. 2c) of any retrieved similar protein, the CE program is invoked. Then the RMSD value is displayed and a new button will appear. By clicking this button, the structural alignment with the query structure is visualized using the Jmol applet on the left panel.
4. *3DZD invariants*: The 121 3DZD invariants of the query are shown in graphical form and in text (Fig. 2g).

3.3 Submit a Batch of Entries

When users want to benchmark 3D-SURFER by submitting many queries, they can use the batch mode page. When submitting a batch of query structures to 3D-SURFER, users can either type a custom list of structure IDs or upload a separate file with those IDs. Then, using the same steps for the submission of a single entry, users can go through the page to select a surface atom representation, a database to compare to, and specify the CATH and the Length filters. The extra step in this section is to select the number of entries to retrieve for each query. Taking 7tim-A as an example, Fig. 3 shows its retrieved results of the top 25 most similar structures.

```

*****
Results for query: 7tim-A
*****
RANK      Structure_ID      EUC_DIST      CATH      LENGTH
*****
1         3ypi-A           1.076         3.20.20.70  247
2         1ypi-B           1.089         3.20.20.70  247
3         3ypi-B           1.175         3.20.20.70  247
4         1ney-A           1.262         3.20.20.70  246
5         2ypi-A           1.284         3.20.20.70  247
6         7tim-B           1.321         3.20.20.70  247
7         1nf0-A           1.333         3.20.20.70  261
8         1ney-B           1.433         3.20.20.70  246
9         1tph-1           1.456         3.20.20.70  245
10        1tph-2           1.492         3.20.20.70  245
11        1tpw-A           1.503         3.20.20.70  245
12        1tpb-1           1.516         3.20.20.70  245
13        1m6j-A           1.522         3.20.20.70  260
14        1tpc-1           1.554         3.20.20.70  245
15        1tpv-A           1.579         3.20.20.70  245
16        1r2r-D           1.604         3.20.20.70  247
17        1tpv-B           1.606         3.20.20.70  245
18        1sw7-A           1.607         3.20.20.70  245
19        1tpw-B           1.610         3.20.20.70  245
20        1sw3-B           1.627         3.20.20.70  247
21        1r2s-D           1.646         3.20.20.70  247
22        1sw0-A           1.651         3.20.20.70  247
23        1tpu-A           1.665         3.20.20.70  245
24        1su5-A           1.678         3.20.20.70  247
25        1tmh-C           1.697         3.20.20.70  254
*****

```

Fig. 3 Retrieved 25 most similar structures for query 7tim-A. EUC_DIST is the Euclidean distance between the query 7tim-A and each retrieved structure. Length shows the number of amino acids residues in the proteins

4 Notes

1. *Surface generation*: The protein surface is constructed using the MSROLL program in Molecular Surface Package [23]. The surface consists of all the points of the van der Waals surface at which a spherical probe (i.e., the solvent radius of 1.4 Å) can touch it. The output of the program is in the form of a polyhedral surface triangulation. Surface computation with MSROLL is failed occasionally. In those cases, the MSMS program (<http://mgltools.scripps.edu/>) is invoked.
2. *Voxelization*: The constructed surface of a protein is mapped on a 3D grid, where voxels (cubic cells) that overlap with the

surface are marked with 1 and 0 otherwise (inside and outside of the protein). This discrete representation of the protein surface is used as an input for computing 3DZD.

3. *Euclidean distance*: The Euclidean distance d_E of two vectors X and Y is defined as:

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

where n is the number of elements in the vectors. n is set to 121 in 3D-SURFER.

4. *Alignment by the CE program*: The CE program is used to align and compute the RMSD of two protein structures. The algorithm first identifies a set of structurally similar main-chain fragment pairs from the two protein structures and then assembles the fragment pairs to form a larger region of similar conformations. Thus, CE compares the main-chain conformations of two protein structures.
5. *Jmol applet*: Jmol is an open-source Java viewer for protein structures in 3D. Jmol can be integrated into web pages to display molecules in various representations. For example, molecules can be displayed as “ball and stick” models, “space filling” models, “ribbon” models, etc. To change the representation of a protein structure, right-click on the Jmol applet area, select the Surface menu, and then the sub-menu Molecular Surface option.
6. *Color of local regions identified by VisGrid or LIGSITE^{esc}*: The size of identified pockets and protrusions is ranked by colors. Red indicates the largest cavity/protrusion/pocket, green indicates the second largest, and blue indicates the third largest cavity/protrusion/pocket. Yellow indicates identified flat regions by VisGrid. The residues in the colored regions are provided in a window (Fig. 2f) on the results page.
7. *Volumes and surface area of a geometric region*: The Qhull program [24] is employed to calculate volumes and the surface area of the geometric regions identified. Qhull calculates volumes and surface areas by computing a convex hull of a specified local surface region (e.g., pocket).
8. *Convex hull*: The convex hull of a set Q of points is the smallest convex polygon P , for which each point in set Q is either on the boundary of P or in its interior. For example, the convex hull of 13 points, p_0 to p_{12} , in a two-dimensional space is given by $CH(Q) = \{p_0, p_1, p_3, p_{10}, p_{12}\}$ in Fig. 4.
9. *The CE alignment file corresponding to a RMSD calculation*: The RMSD between a retrieved structure and a query structure can be computed by checking the RMSD checkbox.

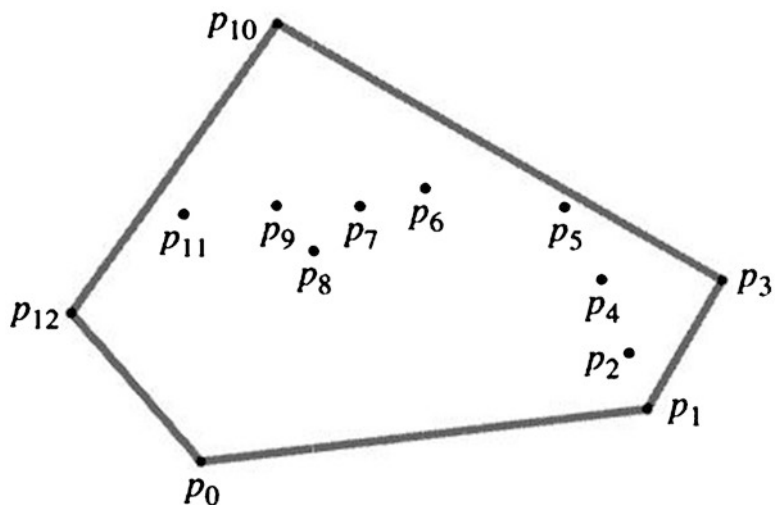


Fig. 4 An example of a convex hull for points in the 2D space

Then, a new button and the RMSD value will be shown. Clicking the RMSD button will invoke Jmol to visualize the structure superimposition between the retrieved structure and the query. The numerical value is a link to a text file that contains the CE alignment result file.

5 Case Studies

We show two examples of 3D-SURFER results. The first example demonstrates that 3DZD detects similar global surface shapes of proteins with a different overall fold and fold class (completely different secondary structure elements) (Fig. 5). In Fig. 5, the protein on the left is a β class protein (PDB code: 1bar-A; fibroblast growth factor), while the one on the right is an α class chain (PDB code: 1rro-A; oncomodulin). The corresponding CATH IDs of the two chains are obviously different, 2.80.10.50 and 1.10.238.10, respectively. Despite the complete difference of the main-chain conformation of the proteins, their surface shapes exhibit surprising similarity with a 3DZD Euclidian distance of 12.66.

The new version of 3D-SURFER extended the coverage of structure databases to include protein domains and complexes, in addition to single chains. A search can be performed between different types of structure data, for example, a single protein chain can be compared against the shape of multi-chain protein complexes. Figure 6 shows a query single chain (PDB code: 2WHH-A) and a complex structure retrieved by a search, which is a homodimeric protein complex (PDB code: 2FDE). A close examination of

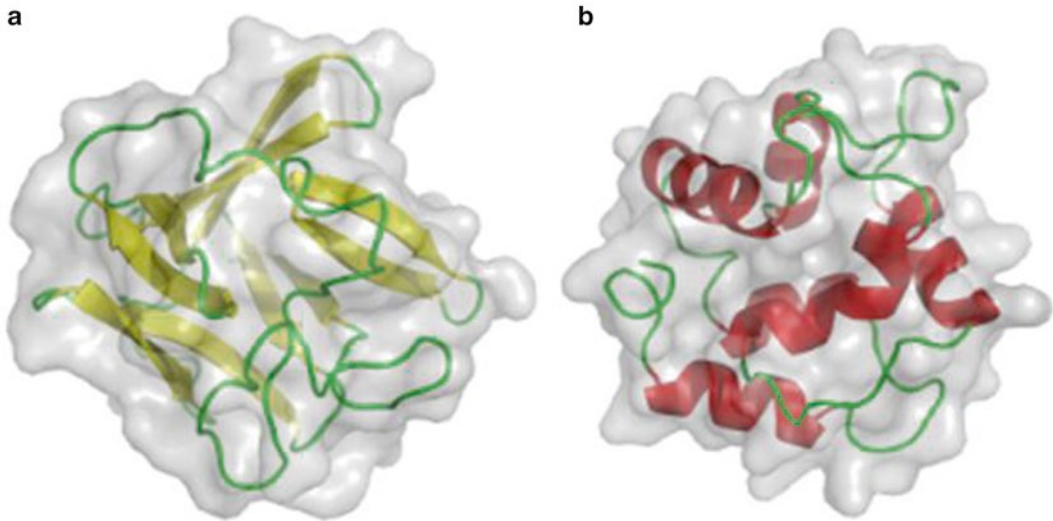


Fig. 5 An example of similar surface shapes of two proteins in different fold class. (a) 1 bar-A, a β -class protein; (b) 1 rro-A, an α -class protein. (The figure is modified from Fig. 5 of ref. [7])

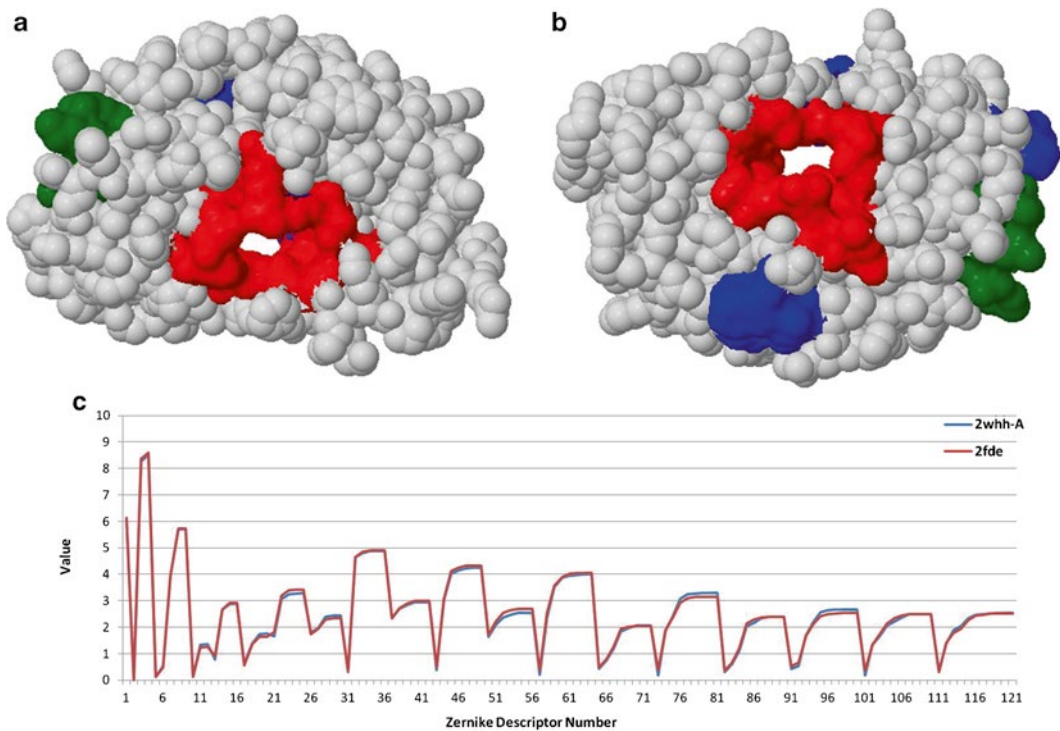


Fig. 6 An example of similar surface shapes of a single protein chain and a two-chain complex. (a) The protein chain 2WHH-A and its largest pocket (*in red color*) identified by LIGSITE^{CSC}; (b) The complex 2FDE, comparable to 2WHH-A both in global shape and in pocket regions; (c) The 3DZD values of 2WHH-A and 2FDE

Fig. 6c shows that these two structures have nearly identical vectors of 3DZD invariants with a Euclidean distance of 0.982. Both of these are HIV-1 proteases [25, 26] with similar enzymatic properties, which are vital targets for the design of antiviral compounds in the treatment of AIDS. 2FDE is a native HIV-1 protease and a homodimeric enzyme in which the active site is located at the subunit interface. 2WHH-A is a mutant HIV-1 protease and a homologous single-chain tethered dimer, which contains a five residue linker, Gly-Gly-Ser-Ser-Gly, that links the N-terminus of the second monomer to the C-terminus of the first monomer. Our results also suggest that the two structures have similar pocket regions identified by LIGSITE^{cs} (shown in Fig. 6a, b). The red regions are pocket sites binding to inhibitors.

Acknowledgments

This work has been supported by grants from the National Institutes of Health (R01GM075004 and R01GM097528), National Science Foundation (EF0850009, DBI1262189, IOS1127027), and National Research Foundation of Korea Grant funded by the Korean Government (NRF-2011-220-C00004). The authors thank Kristen Johnson for proofreading the manuscript.

References

1. Sael L, Chitale M, Kihara D (2012) Structure- and sequence-based function prediction for non-homologous proteins. *J Struct Funct Genomics* 13(2):111–123
2. Hawkins T, Kihara D (2007) Function prediction of uncharacterized proteins. *J Bioinform Comput Biol* 5(1):1–30
3. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389–3402
4. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res* 28(1):235–242
5. Yang JM, Tung CH (2006) Protein structure database search and evolutionary classification. *Nucleic Acids Res* 34(13):3646–3659
6. Yeh JS, Chen DY, Chen BY, Ouhyoung M (2005) A web-based three-dimensional protein retrieval system by matching visual similarity. *Bioinformatics* 21(13):3056–3057
7. 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
8. Kihara D, Sael L, Chikhi R, Esquivel-Rodriguez J (2011) Molecular surface representation using 3D Zernike descriptors for protein shape comparison and docking. *Curr Protein Pept Sci* 12(6):520–530
9. Sael L, Kihara D (2012) Detecting local ligand-binding site similarity in nonhomologous proteins by surface patch comparison. *Proteins* 80(4):1177–1195
10. Chikhi R, Sael L, Kihara D (2011) Protein binding ligand prediction using moments-based methods. *Protein function prediction for Omics Era*, Springer, New York
11. Sael L, Kihara D (2010) Improved protein surface comparison and application to low-resolution protein structure data. *BMC Bioinformatics* 11 Suppl 11:S2
12. Venkatraman V, Chakravarthy PR, Kihara D (2009) Application of 3D Zernike descriptors to shape-based ligand similarity searching. *J Cheminform* 1:19
13. Venkatraman V, Yang YD, Sael L, Kihara D (2009) Protein-protein docking using region-based 3D Zernike descriptors. *BMC Bioinformatics* 10:407

14. Esquivel-Rodriguez J, Yang YD, Kihara D (2012) Multi-LZerD: multiple protein docking for asymmetric complexes. *Proteins* 80(7):1818–1833
15. La D, Esquivel-Rodriguez J, Venkatraman V, Li B, Sael L, Ueng S, Ahrendt S, Kihara D (2009) 3D-SURFER: software for high-throughput protein surface comparison and analysis. *Bioinformatics* 25(21):2843–2844
16. 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
17. Huang B, Schroeder M (2006) LIGSITE^{cc}: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct Biol* 6:19
18. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM (1997) CATH—a hierarchic classification of protein domain structures. *Structure* 5(8):1093–1108
19. Shindyalov IN, Bourne PE (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng* 11(9):739–747
20. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, Murzin AG (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res* 36(Database issue):D419–D425
21. Connolly ML (1983) Solvent-accessible surfaces of proteins and nucleic acids. *Science* 221(4612):709–713
22. Novotni M, Klein R (2003) 3D Zernike descriptors for content based shape retrieval. Paper presented at the proceedings of the 8th ACM symposium on solid modeling and applications, Seattle
23. Connolly ML (1993) The molecular surface package. *J Mol Graph* 11(2):139–141
24. Barber CB, Dobkin DP, Huhdanpaa H (1996) The Quickhull algorithm for convex hulls. *ACM T Math Software* 22(4):469–483
25. Prashar V, Bihani S, Das A, Ferrer JL, Hosur M (2009) Catalytic water co-existing with a product peptide in the active site of HIV-1 protease revealed by X-ray structure analysis. *PLoS One* 4(11):e7860
26. Miller JF, Andrews CW, Brieger M, Furfine ES, Hale MR, Hanlon MH, Hazen RJ, Kaldor I, McLean EW, Reynolds D, Sammond DM, Spaltenstein A, Tung R, Turner EM, Xu RX, Sherrill RG (2006) Ultra-potent P1 modified arylsulfonamide HIV protease inhibitors: the discovery of GW0385. *Bioorg Med Chem Lett* 16(7):1788–1794