Structural bioinformatics

3D-SURFER: software for high-throughput protein surface comparison and analysis

David La1,†, Juan Esquivel-Rodriguez2,†, Vishwesh Venkatraman1, Bin Li2, Lee Sael2, Stephen Ueng2, Steven Ahrendt1 and Daisuke Kihara1,2,3,∗

1Department of Biological Sciences, 2Department of Computer Science and 3Markey Center for Structural Biology, College of Science, Purdue University, West Lafayette, IN 47907, USA

Received on June 12, 2009; revised on July 24, 2009; accepted on September 10, 2009
Advance Access publication September 16, 2009
Associate Editor: Anna Tramontano

ABSTRACT

Summary: We present 3D-SURFER, a web-based tool designed to facilitate high-throughput comparison and characterization of proteins based on their surface shape. As each protein is effectively represented by a vector of 3D Zernike descriptors, comparison times for a query protein against the entire PDB take, on an average, only a couple of seconds. The web interface has been designed to be as interactive as possible with displays showing animated protein rotations, CATH codes and structural alignments using the CE program. In addition, geometrically interesting local features of the protein surface, such as pockets that often correspond to ligand binding sites as well as protrusions and flat regions can also be identified and visualized.

Availability: 3D-SURFER is a web application that can be freely accessed from: http://dragon.bio.purdue.edu/3d-surf.

Contact: dkihara@purdue.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Greater insight into the inner workings of the cellular machinery will become more critical as the current structural genomics initiatives for solving protein structures at high-throughput rates continue to rapidly progress. Given the huge number of experimentally solved but largely uncharacterized structures, the understanding of protein biochemical function assumes great importance. Although numerous representations of proteins have been used, surface-based approaches have been found to be quite useful both by way of analysis and visualization (Venkatraman et al., 2009, Fischer et al., 1993; Rosen et al., 1998).

Traditional protein structure comparison techniques make use of the pairwise alignment of protein Cβ backbone or all atom structure representations. However, computing alignments has a high time complexity and is unsuitable for applications such as real-time structure database searches. To obviate this difficulty, methods such as 3D-BLAST encode the structure as a 1D sequence of alphabets (Yang and Tung, 2006; Supplementary Material). Light Field Descriptors, on the other hand, create 2D projections (combination of 2D Zernike and Fourier coefficients) rendered from uniformly distributed points around a sphere that surrounds the protein (Yeh et al., 2005). More recently, the development of 3D moment-based shape representations have shown promising performance for large-scale comparisons (Bustos et al., 2007). Among these, the 3D Zernike descriptors (3DZD) have been found to be suitable for the efficient comparison of protein surfaces (Sael et al., 2008). Unlike the previous two methods, which compare 1D or 2D representations, 3DZD are based on a 3D function expansion.

Here, we present 3D-SURFER, a web-based environment for high-throughput protein surface comparison and analysis. The server compares a single protein surface against all protein structures in PDB in just a couple of seconds (over 130 000 single chain structures, more than 55 000 total PDB entries, which are updated monthly). A performance comparison against other similar tools can be found in the previous work (Sael et al., 2008). In addition, local geometrical characteristics of a protein, which represent potential ligand binding sites, can be identified by the VisGrid algorithm (Li et al., 2008). Results shown include visual aids in the form of animated rotations of proteins along with the associated CATH codes (Orengo et al., 2002), and structure alignment calculations using the Combinatorial Extension (CE) algorithm (Shindyalov and Bourne 1998).

2 METHODS

2.1 3D Zernike descriptors

3DZD are utilized for the efficient comparison of protein surfaces across the entire PDB. The calculation of the invariants starts by voxelizing the protein molecular surface that is triangulated by MSROLL version 3.9.3 (Connolly, 1993). The mesh is then discretized to generate a cubic grid. 3DZD, a vector of 121 numbers, is then computed for the protein surface represented by the grid voxels. A single protein represented as a vector can be compared with other structures simply using the Euclidean distance. An example of the retrieval by 3DZD is shown in Table 1.

We have shown in our previous work that structure retrieval by 3DZD agrees well with main-chain comparison by CE (Sael et al., 2008). Also it was found that surface comparison by 3DZD can identify functionally related proteins that cannot be discovered otherwise, due to distant evolutionary relationship (Sael et al., 2008).
A protein can be interactively analyzed by VisGrid, which identifies geometric features of protein surfaces, i.e. pockets, protrusions, hollow spaces and flat regions (Li et al., 2008), which are often associated with binding sites. VisGrid uses a novel visibility criterion, which essentially indicates the fraction of open directions for a given point on the protein surface. The three largest pockets and protrusions are reported. The Qhull program (Barber et al., 1996) is used to calculate volumes and surface area of the pockets identified.

### Table 1. Top 10 results using the query 2mta-A

<table>
<thead>
<tr>
<th>Rank</th>
<th>PDB ID</th>
<th>CATH Code</th>
<th>Euclidean Distance</th>
<th>RMSD (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1t5k-A</td>
<td>2.60-40.420</td>
<td>1.575</td>
<td>0.48</td>
</tr>
<tr>
<td>2</td>
<td>2id-A</td>
<td>2.60-40.420</td>
<td>1.643</td>
<td>0.44</td>
</tr>
<tr>
<td>3</td>
<td>1t5k-B</td>
<td>2.60-40.420</td>
<td>1.777</td>
<td>0.77</td>
</tr>
<tr>
<td>4</td>
<td>1t5k-C</td>
<td>2.60-40.420</td>
<td>1.778</td>
<td>0.38</td>
</tr>
<tr>
<td>5</td>
<td>1aan-A</td>
<td>2.60-40.420</td>
<td>1.840</td>
<td>0.55</td>
</tr>
<tr>
<td>6</td>
<td>2dta-A</td>
<td>2.60-40.420</td>
<td>2.160</td>
<td>0.89</td>
</tr>
<tr>
<td>7</td>
<td>1d2-C</td>
<td>2.60-40.420</td>
<td>2.201</td>
<td>0.60</td>
</tr>
<tr>
<td>8</td>
<td>1aa-A</td>
<td>2.60-40.420</td>
<td>2.215</td>
<td>0.53</td>
</tr>
<tr>
<td>9</td>
<td>1d2-A</td>
<td>2.60-40.420</td>
<td>2.320</td>
<td>0.56</td>
</tr>
<tr>
<td>10</td>
<td>2dta-A</td>
<td>2.60-40.420</td>
<td>2.322</td>
<td>1.00</td>
</tr>
</tbody>
</table>

*aStructural alignments calculated using CE*

### 2.2 Analyzing local surface geometry

A protein can be interactively analyzed by VisGrid, which identifies geometric features of protein surfaces, i.e. pockets, protrusions, hollow spaces and flat regions (Li et al., 2008), which are often associated with binding sites. VisGrid uses a novel visibility criterion, which essentially indicates the fraction of open directions for a given point on the protein surface. The three largest pockets and protrusions are reported. The Qhull program (Barber et al., 1996) is used to calculate volumes and surface area of the pockets identified.

### 2.3 Input

3D-SURFER takes a PDB ID as an input structure to compare against the entire PDB. PDB IDs may be followed by a character representing the chain. For example, if the PDB structure 2MTA and chain A is of interest, the text entry should be 2MTA-A. Alternatively, a custom PDB structure may be uploaded and utilized as the query. In either case, a search against the entire structure database is executed on-the-fly. Additionally, the user can specify two types of filtering: CATH filtering that avoids displaying structures with similar CATH levels, and length filtering, in charge of displaying proteins whose lengths are similar to the query structure.

### 2.4 Output

The right section of the results panel lists the structures identified as similar by clicking on the buttons called Cavity, Protrusion or Flat. The interface will render the surface in three different colors based on their rank in terms of geometric visibility: Red (1st), Green (2nd) and Blue (3rd). The volumes will render the surface in three different colors based on their rank in terms of geometric visibility: Red (1st), Green (2nd) and Blue (3rd). The volumes and surface areas of each region are also shown.

### 3 SUMMARY

3D-SURFER provides a platform to perform both global and local structure analysis in real time. Similarity in global structure infers evolutionary relationship in many cases, which can give a clue for the function of the protein. We plan to incorporate protein pocket database search into our platform in the future. In addition, protein surface properties such as electrostatic potentials, hydrophobicity and conservation will be integrated into 3D-SURFER for detailed analysis designed to assist investigating function of proteins.

### 4 SUPPORTED PLATFORMS

All latest web browsers are supported. The Java plug-in, and appropriate configuration, is required for visualization using Jmol.

**Funding:** National Institutes of Health (R01 GM075004); National Science Foundation (DMS 0604776, DMS0800568).

**Conflict of Interest:** none declared.

### REFERENCES


