SURF'S UP! – Protein classification by surface comparisons

Joanna M Sasin[1,*], Adam Godzik[2] and Janusz M Bujnicki[1]

[1]*Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology, Warsaw, Poland*

[2]*The Burnham Institute, La Jolla, CA, USA*

**Corresponding author (Email, asia@genesilico.pl)*

Large-scale genome sequencing and structural genomics projects generate numerous sequences and structures for 'hypothetical' proteins without functional characterizations. Detection of homology to experimentally characterized proteins can provide functional clues, but the accuracy of homology-based predictions is limited by the paucity of tools for quantitative comparison of diverging residues responsible for the functional divergence. SURF'S UP! is a web server for analysis of functional relationships in protein families, as inferred from protein surface maps comparison according to the algorithm. It assigns a numerical score to the similarity between patterns of physicochemical features (charge, hydrophobicity) on compared protein surfaces. It allows recognizing clusters of proteins that have similar surfaces, hence presumably similar functions. The server takes as an input a set of protein coordinates and returns files with "spherical coordinates" of proteins in a PDB format and their graphical presentation, a matrix with values of mutual similarities between the surfaces, and the unrooted tree that represents the clustering of similar surfaces, calculated by the neighbor-joining method. SURF'S UP! facilitates the comparative analysis of physicochemical features of the surface, which are the key determinants of the protein function. By concentrating on coarse surface features, SURF'S UP! can work with models obtained from comparative modelling. Although it is designed to analyse the conservation among homologs, it can also be used to compare surfaces of non-homologous proteins with different three-dimensional folds, as long as a functionally meaningful structural superposition is supplied by the user. Another valuable characteristic of our method is the lack of initial assumptions about the functional features to be compared. SURF'S UP! is freely available for academic researchers at *http://asia.genesilico.pl/surfs_up/*.

[Sasin J M, Godzik A and Bujnicki J M 2007 SURF'S UP! – Protein classification by surface comparisons; *J. Biosci*. **32** 97–100]

1. Introduction

With an increasing number of experimentally uncharacterized protein sequences and structures produced by genome sequencing or structural genomic initiatives, we often encounter large protein families with only a few members of known function. Prediction of function for a protein by simple detection of homology to other, well-characterized proteins, can provide useful hints, but usually can be confident only on a general level e.g. the type of the reaction catalyzed by an enzyme, but not its specificity. In particular, multiple paralogs with similar functions but different specificities are often present in a genome and it is very difficult to predict the level of functional specialization only by comparing their amino acid sequences. Frequently, it is challenging even to confidently answer the question whether the detailed function of a particular protein is conserved compared to its homolog for which the structure is available (reviews: Rost *et al* 2003; Friedberg 2006). Although comparative modelling can provide us with models of members of the entire family and the detailed analysis of such models can provide insights as to the possible functional differences between various members of the family, there is no simple method that would allow us to quantify the distribution of features relevant for the functional divergence in the entire family.

The protein function is determined mostly by residues forming the interaction sites – specific groups forming ligand-binding and catalytic pockets or hydrophobic patches

Keywords. Protein structures; protein surfaces; structural bioinformatics

defining protein-protein binding sites – most if not all of them located on the surface of the protein. Comparing surfaces of proteins provides a first hint about the possible similarities or differences between their functions. In fact, visual comparison of surface features of homologous proteins is often used in the scientific literature to illustrate the molecular basis of their functional similarities or differences. Here we present a tool that facilitates such analyses by automating the calculation, quantitative comparison, and visualization of protein surface maps.

There are several methods that analyse or compare protein surfaces. However, they either concentrate on some local features specific to a small region of the protein surface, such as protein active sites or binding cavities or surface patches that may be involved in protein-protein interactions (Kinoshita and Kamura 2003). The obvious reason for the local character of such algorithms is that protein surfaces are very irregular and therefore difficult to compare. The algorithm implemented in SURF'S UP!, which we refer to as molecular cartography (Pawlowski and Godzik 2001), circumvents this problem by approximating a protein surface as a sphere, which allows for easy visualization and comparison of distributions of physico-chemical features of the surface, such as a charge, polarity and hydrophobocity.

2. Methods

SURF'S UP! is a WWW-based server that accepts experimentally solved or modelled protein structures and uses a molecular cartography approach to perform a pairwise comparison of all models. The features of the protein surface to be compared include polarity, hydrophobocity and charge of the amino acids residues. The coordinates of structures to be compared must be provided in a multiple structure file in a PDB-format. The algorithm does not perform any superposition or rotation of the models, thus it is strongly recommended that the user submits structures that are already superimposed structures. The input file can be obtained by manual matching of structures by the SwissPDBViewer program (*http://www.expasy.ch/spdbv*) (Guex and Peitsch 1997) or, if the user is confident in the quality of automated structure alignments, one of the tools available online, such as Multiprot (*http://bioinfo3d.cs.tau.ac.il/MultiProt/*). If the structures to be analysed are not superimposed (or if the user wishes to compare the surfaces of protein structures, whose backbone coordinates are not superimposable, e.g. non-homologous), SURF'S UP! provides an option to fit the surfaces.

We must emphasize that fitting of surfaces without superimposing the protein backbones is risky and the results must be interpreted with caution, as artifacts may be generated (especially if the surfaces outside the functionally important regions are highly divergent). Even for comparison of non-homologous enzymes, we recommend the submission of structures with the key active site residues superimposed, so the server could be used to identify similarities of substrate-binding pockets around the active sites, regardless of homology.

Based on the results of pairwise comparisons, provided that at least three structures were included in the input file, SURF'S UP! calculates an unrooted tree using the neighbor-joining algorithm (Godzik *et al* 1992), to visualize the clustering of most similar surfaces. The protein surfaces are approximated by spheres, centered on the protein's center of mass. For each model, the amino acid properties of surface residues are projected on the sphere for easy visualization and comparison. Only heavy (non-hydrogen) atoms of exposed residues are considered (there is also an option to use only C-beta atoms of site-chains, which is recommended for low-resolution structures, e.g. homology models). Solvent accessibility is calculated as reported earlier (Pawlowski and Godzik 2001) and the exposed residues are defined as those with more than 40% solvent accessibility. The resulting spherical maps of amino acid features are compared and the comparison result is expressed as a single value, the map 'distance'. For this purpose the amino acid properties are projected on the regular grid. The surface grid points are chosen, starting from t=0 (the 'equator'). The distance between the equator grid points is called a map diffusion parameter. By default this value is set at 15 degrees. Amino acid residues are divided into four classes, depending on their chemical features: hydrophobic (I, V, L, A, F, M, W, P, C), acidic (D, E), basic (H, K, R) and polar (N, Q, T, S, G, Y). For each of the areas, limited by lines connecting the grid points, the number of the amino acids belonging to one or more of the classes (it depends on the user choice) is counted to give a density of a given feature in a given grid square. The similarity between corresponding areas in two proteins is counted as an Euclidean distance between two points in four dimensional space, whose coordinates describe the estimated densities of amino acids belonging to each class. The mean value from scores obtained for all the areas is calculated. Finally, the results are rescaled between 1 (highest similarity score) and 0 (lowest similarity observed). Pairwise comparisons are made for each pair of the proteins to be analysed. The results of the pairwise comparison are expressed as a similarity matrix stored in the PHYLIP format. SURF'S UP! implements the Neighbor-Joining method of Saitou and Nei (1987) to create the unrooted tree (a cladogram of surface similarities), which is exported in the NEWICK format. Additionally, colour images of protein surfaces are made using MOLSCRIPT (Kraulis 1991).

When the uploaded structures are not superimposed, the spherical maps of proteins are 'superimposed' while comparison procedure. For each pair of proteins the first map is chosen as a reference one. The second sphere is

rotated around equator and meridian. The similarity between maps in each possible position is evaluated. The best score for each pair is taken as a final result.

The whole procedure, from sending an input file to getting results, should take below 1 min for up to 10 superimposed structures. In our tests, for the sets of as many as 50 superimposed structures, the execution time has not exceeded 2 min. If protein structures in the input file are indicated as not superimposed, the execution time will differ, depending on the protein size and the complexity of the structure. SURF'S UP currently runs on a general purpose server with an Intel Pentium III processor. We are planning deploying it on a 64-bit SUN machine dedicated to the www services, which should improve the execution time.

3. Discussion

SURF'S UP! is a web-server implementation of the protein surface comparison methods originally described by Pawlowski and Godzik (2001). It was tested on several different protein families, including the CARD/DD/DED domain superfamily (Rost *et al* 2003; Guex and Peitsch 1997) and the protein tyrosine phosphatase family. The results, e.g. clustering of proteins with similar functions and clear separation of protein with (potentially) different function agree with the classifications obtained by manual analyses. Figure 1 shows result of analysis for members of protein tyrosine phosphatases family. Proteins that belong to the dual specificity PTPs and Tyr phosphatases group in separate clusters.

It must be emphasized that the approximations made in the algorithm do not allow for detailed analyses of topographic, electrostatic and hydrophobic properties of protein surfaces. However such studies are feasible only when high-resolution structures are compared, while SURF'S UP! has been designed aiming mainly at comparison of mixed data sets, including both protein structures determined by crystallography and NMR and computational models of low to moderate accuracy, e.g. obtained by the fold-recognition methods via our structure prediction Meta-Server (Kurowski and Bujnicki 2003). Another valuable characteristic of the method presented here is the lack of initial assumptions about the character of the functional differences among the proteins to be compared.

Our method has several shortcomings that have to be mentioned: First, the spherical representation is appropriate only for isolated, globular domains, and in the analyses of multi-domain or elongated proteins severe distortion may be introduced. Thus, as a minimum, we recommend splitting the proteins to be analysed into individual domains.

The analysis of surfaces by SURF'S UP! can serve as a complement to a number of other methods for prediction of protein function that typically address the global similarity of sequences and structures (mostly conserved residues in the protein core) or search for pockets, clefts or constellations of a few functionally important residues (Binkowski *et al*. 2004). While these methods are better suited for the prediction of general function (global similarity) or identification of very precise features (identification of catalytic residues), SURF'S UP! focuses on the conservation of intermediate features that are typically associated with the specificity of interactions e.g. a preference for a particular type of macromolecular binding partner that interacts with a relatively large region of protein surface.

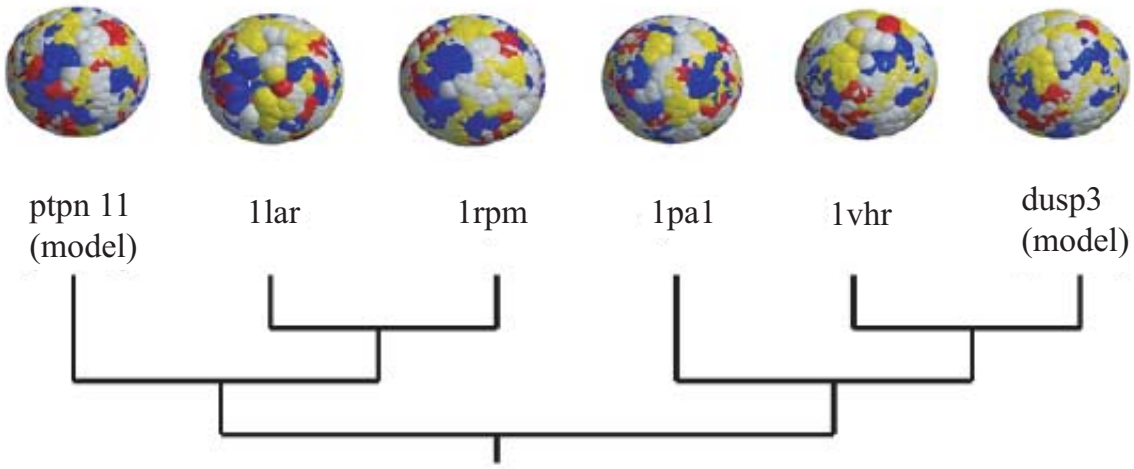

Figure 1 . An example of the analysis carried out using SURF'S UP!. Members of the protein tyrosine phosphatase (PTP) family were analysed. Our program have correctly grouped together dual specificity phosphatases (the cluster including experimentally determined structures 1pa1 and 1vhr, and a comparative model of DUSP3) to the exclusion of the phosphatase Y family (PDB structures 1lar, 1rpm and a comparative model of phosphatase Y). The images of protein surfaces approximated as spheres are colour coded: hydrophobic residues in grey, acidic residues in red, hydrophilic residues in yellow and basic residues in blue.

In the case of comparison of homologous proteins, we recommend the use of SURF'S UP! together with the STRUCLA server developed in our group (Sasin *et al*. 2003) that infers trees based on comparison of the backbone coordinates, without any consideration of surfaces or amino acid residue similarity. Identification of inconsistencies between the fold-based and surface-based trees can suggest interesting shifts in the evolutionary rates, leading to important functional predictions, but it can also highlight artifacts caused by the spherical representation of the protein surface. In the future, we plan to modify SURF'S UP as to recognize differences in the protein shape and warn the user about the potential artifacts.

Acknowledgements

This analysis was funded by the MEiN (grant KBN–1581/T11/2005/29). JMB was supported by the grant PBZ–KBN–088/PO4/2003.

References

Binkowski T A, Freeman P and Liang J 2004 pvSOAR: detecting similar surface patterns of pocket and void surfaces of amino amid residues on proteins; *NAR* **32** W555–W558

Friedberg I 2006 Automated protein function prediction–the genomic challenge; *Brief Bioinform.* **7** 225–242

Godzik A, Kolinski A and Skolnick J 1992 A topology fingerprint approach to the inverse folding problem; *J. Mol. Biol.* **227** 227–238

Guex N and Peitsch M C 1997 Swiss-model and the Swiss-PdbViewer: An environment for comparative protein Modeling; *Electrophoresis* **18** 2714–2723

Kinoshita K and Kamura H 2003 Identification of protein biochemical functions by similarity search using the molecular surface database eF-site; *Protein Sci.* **12** 1589–1595

Kraulis J 1991 MOLSCRIPT: A Program to Produce Both Detailed and Schematic Plots of Protein Structures; *J. Appl. Crystallogr*. **24** 946–950

Kurowski M A and Bujnicki J M 2003 GeneSilico protein structure prediction meta-server; *Nucleic Acids Res.* **31** 3305–3307

Pawlowski K and Godzik A 2001 Surface Map Comparison: Studying Function Diversity of Homologous Proteins; *J. Mol. Biol.* **309** 793–800

Rost B, Liu J, Nair R, Wrzeszczynski K O and Ofran Y 2003 Automatic prediction of protein function; *Cell. Mol. Life Sci.* **60** 2637–2650

Saitou N and Nei M 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees; *Mol. Biol. Evol.* **4** 406–425

Sasin J M, Kurowski M A and Bujnicki J M 2003 STRUCLA: a WWW meta-server for protein structure comparison and evolutionary classification; *Bioinformatics (Suppl. 1)* **19** 252–254

ePublication: 12 December 2006