SURFMAP manual

Hugo Schweke

February 8, 2022

Contents

I - What is SURFMAP?

In a nutshell, SURFMAP is a small software designed to compute the surface properties of a protein, and to map this property on a 2-D plan through a sinusoidal projection. It enables the rapid visualization of a surface property across the whole surface of a protein. Such view can prove to be very useful when comparing the surface properties of homologous proteins, for example.

II - A basic overview

Surface properties mapped by SURFMAP

The properties that SURFMAP is mapping must be chosen among:

- stickiness (1): measure of the propensity of an amino acid to be enriched (or depleted) at protein binding sites. The stickiness enables to detect regions that are theoretically more prone to interaction, i.e. "more sticky", from other regions.
- Kyte-Doolittle hydrophobicity (2): measure of the degree of hydrophobicity/hydrophilicity of amino acids according to the Kyte-Doolittle hydrophobicity scale.
- Wimley–White hydrophobicity (3): measure of the degree of hydrophobicity/hydrophilicity of amino acids according to the Wimley–White hydrophobicity scale.
- circular variance (4): The circular variance (CV) measures the atom density around each atom. The CV provides a useful descriptor of the local geometry of a surface region. CV values range between 0 and 1 with low values reflecting protruding residues, and high values indicating residues located in cavities. This measure can be calculated at the residue level (mean CV of the residue's atoms), or at the atom level.
- electrostatic potential: SURFMAP can call the APBS software (5) that will compute the electrostatic potential, using the CHARMM forcefield (6), and map the resulting potential on a 2-D map.
- b-factor: the user have the possibility to map any value contained in the b-factor of the pdb structure provided in input.
- binding sites: this option is similar to the b-factor option, but use a discrete set of colors, instead of a continuous one. When using this option, please provide a pdb file with discrete values in b-factor.

Quick usage of SURFMAP

SURFMAP needs in input a pdb file. If you installed the dockerized version of SURFMAP and created an alias `surfmap`, you will simply need to type the alias in the terminal followed by the required arguments. If you are not working with the dockerized version of SURFMAP, you can launch it by executing the script `surfmap.py`. This script serves as a pipeline and will launch successively all the necessary scripts.

In the directory example there are several files (a pdb file, and 2 coordinates mapping file) to familiarize yourself with SURFMAP.

Basically there are two mandatory arguments: the input pdb file, and the surface property you want to map. All the rest is optional. In the next section you will see more in details how to use SURFMAP.

Usage of SURFMAP

In the following section, we will assume that the SURFMAP program is called through the `surfmap` command that is either an alias of `run_surfmap_image.py` as shown here or directly an alias `surfmap.py` if you have installed SURFMAP locally on your machine.

To guide the user on how to use SURFMAP, we will use files in the `example/` directory that can be found in the downloaded SURFMAP project. The files in the directory are:

- `1g3n_A_CV.pdb`
- `1g3n_A.pdb`
- `1gv3_A_binding_sites.pdb`
- `example_1g3n_A_stickiness/`
- `example_1g3n_A_stickiness_allfiles/`
- `example_1g3n_A_stickiness_mapping_residues/`
- `example_1gv3_A_stickiness_binding_sites/`
- `example_mapping_stickiness/`
- `README`
- `residues_to_map.txt`

SURFMAP inputs and outputs

SURFMAP allows to compute different protein surface features and to map them on a 2-D plan through a sinusoidal projection. Thus two mandatory arguments must be given as inputs by the user: `-pdb` and `-tomap`:

- the `-pdb` argument must be followed by the protein structure in PDB format the user wants to analyse
- the `-tomap` argument must be given a keyword representing the protein surface feature the user wants to map. The user can also use the option `all` to map the Kyte-Doolittle hydrophobicity, the Wimley-White hydrophobicity, the stickiness and the circular variance per residue at the same time. The available keywords are listed below (see SURFMAP_manual.pdf or the original article for a description):
  - wimley_white
  - kyte_doolittle
  - stickiness
  - circular_variance
  - circular_variance_atom
  - electrostatics
  - bfactor
  - binding_sites
  - all

For instance, the following command line will map the stickiness protein surface feature for the chain A of the protein 1G3N:

```
surfmap -pdb 1g3n_A.pdb -tomap stickiness
```

The above command will generate an output directory named `output_SURFMAP_1g3n_A_stickiness/` with the following content:

- `log_parameters`: a summary of the basic parameters used to compute the map
- `maps/1g3n_A_stickiness_map.pdf`: the generated 2-D map in pdf format
- `smoothed_matrices/1g3n_A_stickiness_smoothed_matrix.txt`: a computed smoothed matrix file used to generate the 2-D map

Here are several SURFMAP commands to calculate maps with different parameters:

SURFMAP command to compute a stickiness map of the pdb file 1g3n_A.pdb. The option `--keep` means that all the intermediary files are kept. Resulting files are in the directory `example_1g3n_A_stickiness_allfiles/`:

```
surfmap -pdb 1g3n_A.pdb -tomap stickiness --keep
```

SURFMAP command to compute maps with residues from file residues_to_map.txt mapped directly on the map. Resulting files are in the directory `example_1g3n_A_stickiness_mapping_residues/`:

```
surfmap -pdb 1g3n_A.pdb -tomap stickiness -res residues_to_map.txt
```

SURFMAP command to compute map of bfactor with discrete values (initially developped for binding site mapping). Resulting files are in the directory `example_1gv3_A_stickiness_binding_sites/`:

```
surfmap -pdb 1gv3_A_binding_sites.pdb -tomap binding_sites
```

Warnings

- SURFMAP does not take into account heteroatoms (HETATM). Please be careful that your pdb file does not contain special residues, such as selenocystein for example. The hydrophobicity and stickiness scales were not designed to take into account these special residues.
- The different protonation states of the HIS amino acid are considered the same as HIS.
- The first step of the cartography involves a call to MSMS (7). So you need to give in input a file that can be handled by MSMS.

III - Complete list of arguments of run_surfmap_image.py / surfmap.py

run_surfmap_image.py / surfmap.py can take the following arguments:

-pdb

Input file. The input file need to be in pdb format.

Rk: The user must either use the `-pdb` option or `-mat` option (see below).

-mat

Input file. In that case the user needs to provide a matrix file, in the same format as SURFMAP matrices files (see format in the examples): | phi | theta | score |

Rk: The user must either use the `-pdb` option or `-mat` option (see above).

-tomap

Surface property of the protein to map There are the following possibilities: stickiness, Kyte-Doolittle hydrophobicity, Wimley-White hydrophobicity, electrostatics potential, circular variance or any value in the bfactor of the input pdb file. If the user chose the option “all”, the stickiness, circular variance, Kyte-Doolittle

hydrophobicity and Wimley-White hydrophibicity are mapped. (**MANDATORY**)

-proj

Projection used to map the protein surface on a 2D plan. SURFMAP can map proteins surface using three different projections: sinusoidal (= Sanson-Flamsteed), Mollweide or Lambert cylindrical. The default choice is sinusoidal. (**OPTIONAL**)

-rad

radius of the atoms used to create the shell. Default value is 3.0Å, normally you do not have to choose another radius. (**OPTIONAL**)

-d

name of the output directory. By default a directory "output_SURFMAP_INPUTPDB_PROPERTY" is created in the current directory (for example output_SURFMAP_1g3n_A_stickiness/). (**OPTIONAL**)

-s

size of a grid cell (in degrees). Default value is 5, which results in a 72 (360/5) x 36 (180/5) grid map. Value must be congruent with 180 (180%%value = 0), otherwise SURFMAP will raise an error message. (**OPTIONAL**)

-res

to map the coordinates of residues on the 2-D map (**OPTIONAL**).
Format is the following: CHAIN RESNB RES
With CHAIN the chain index in the input pdb, RESNB the residue number in the input pdb, and RES the residue type.
Important: residue numbers must correspond to the numbering in the input pdb file.

--nosmooth

With this option the smoothing step is skipped (By default output maps are smoothed). (**OPTIONAL**)

--png

option to generate a png file. By default SURFMAP generates a pdf file. (**OPTIONAL**)

--keep

option to keep all intermediary files. By default SURFMAP keeps only a pdf + txt file of the smoothed matrix. (**OPTIONAL**)

IV - List of scripts

Here are detailed the scripts and modules that compose SURFMAP.

Main directory

- run_surfmap_image.py
It is a python script that acts as a proxy to call the script surfmap.py through the SURFMAP docker image. It makes invisible to the user complex command lines required to create a bridge between the host filesystem and the container filesystem that are isolated by default.

- surfmap.py
Launcher script for SURFMAP. Type "python surfmap.py -h" to see all the arguments. This is this script that will execute all the other SURFMAP's scripts.

Directory "scripts"

- computeCoordList.R computes coord_list files from partlist.out files

a coord_list file is is a 6 columns file listing, for each cell grid of the map, all the residues included. the format is the following:

absc | ord | score | resnb | restype | chain | index_sol

with:
-absc: abscissa coordinate on the map
-ord: odinate coordinate on the map
-score: score of the property tested
-resnb: residue number in the input pdb
-restype: residue type
-chain: residue chain
-index_sol: index of the generated particle

- computeMaps.R
script that computes a pdf map from a smoothed matrix file

- computeMatrices.R
script that computes a smoothed matrix from a coordlist file. The format is the following:

absc | ord | value | residues

with:
-absc: abscissa coordinate on the map
-ord: ordinate coordinate on the map
-value: the value of the property tested. A value of 100 means that the grid cell is outside the projection. A value of 'NA' means that there is no particles in the grid cell.
-residues: list of residues "belonging" to the couple of coordinates (= to the grid cell). The format is RESTYPE_RESNB_CHAINID. If several residues are present, they are separated by a comma.

- compute_shell.sh
This script is mainly use to call MSMS to compute the solvent excluded surface of the protein, from which the "shell" (i.e. the point-particles at the surface of the protein) is derived. It also calls APBS to compute electrostatic potential if the user requests to map electrostatic potential.

- multival_csv_to_pdb.py
script used specifically for electrostatics calculation. It simply modifies the format of the output file of the APBS executable “multivalue”.

Directory “tools”

- Structure.py
Contains a set functions used by SURFMAP

- SurfmapTools.py
script that associate the closest residues to each particles. compute all the surface properties values.

V - References

(1) Levy ED, De S, Teichmann SA. Cellular crowding imposes global constraints on the chemistry and evolution of proteomes. Proc Natl Acad Sci U S A. 2012 Dec 11;109(50):20461-6. doi: 10.1073/pnas.1209312109. Epub 2012 Nov 26. PMID: 23184996; PMCID: PMC3528536.

(2) Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol. 1982 May 5;157(1):105-32. doi: 10.1016/0022-2836(82)90515-0. PMID: 7108955.

(3) Wimley WC, White SH. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. Nat Struct Biol. 1996 Oct;3(10):842-8. doi: 10.1038/nsb1096-842. PMID: 8836100.

(4) Mezei M. A new method for mapping macromolecular topography. J Mol Graph Model. 2003 Mar;21(5):463-72. doi: 10.1016/s1093-3263(02)00203-6. PMID: 12543141.

(5) Jurrus E, Engel D, Star K, Monson K, Brandi J, Felberg LE, Brookes DH, Wilson L, Chen J, Liles K, Chun M, Li P, Gohara DW, Dolinsky T, Konecny R, Koes DR, Nielsen JE, Head-Gordon T, Geng W, Krasny R, Wei GW, Holst MJ, McCammon JA, Baker NA. Improvements to the APBS biomolecular solvation software suite. Protein Sci. 2018 Jan;27(1):112-128. doi: 10.1002/pro.3280. Epub 2017 Oct 24. PMID: 28836357; PMCID: PMC5734301.

(6) Mackerell AD Jr, Feig M, Brooks CL 3rd. Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. J Comput Chem. 2004 Aug;25(11):1400-15. doi: 10.1002/jcc.20065. PMID: 15185334.

(7) Sanner MF, Olson AJ, Spehner JC. Reduced surface: an efficient way to compute molecular surfaces. Biopolymers. 1996 Mar;38(3):305-20. doi: 10.1002/(SICI)1097-0282(199603)38:3%3C305::AID-BIP4%3E3.0.CO;2-Y. PMID: 8906967.