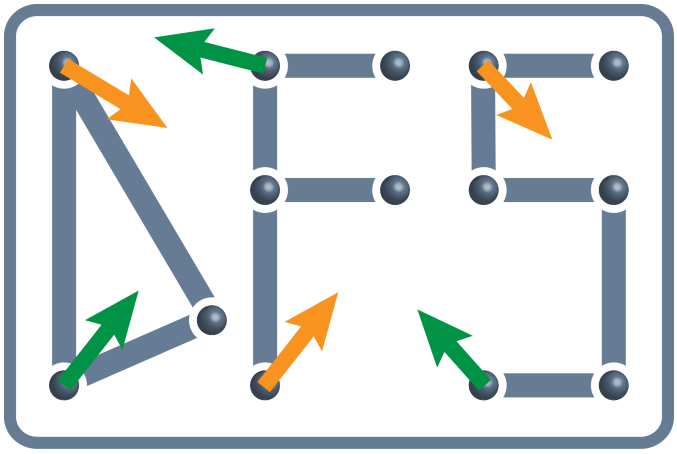

Double Force Scanning

A user-friendly Python implementation

Matteo Tiberti

https://fornililab.github.io/dfs/

1 Introduction

This guide introduces each script in the dfs package, with detailed description of the available options and of the expected outputs.

2 Scripts

2.1 dfs

The dfs script is the main script of the collection. Its purpose is to use the DFS method as described in [M. Tiberti, A. Pandini, F. Fraternali, A. Fornili, "In silico identification of rescue sites by double force scanning", submitted] to generate matrices of rescuability scores and accessory data. The method uses forces applied through Linear Response Theory to calculate a rescuability score *S* between (first site, second site) pairs of residues. By default, every pair of residues is tested and the outcome is a *N* x *N* matrix of rescuability scores S_{ij}, where *N* is the number of residues. A positive value of S_{ij} indicates that mutations at the second site *j* can have a compensatory effect on mutations at the first site *i*. Running with the default options, dfs just requires a single PDB input file with option –p, e.g. dfs –p 1TSR.pdb.
The different groups of options will be discussed in the following.

General options

-p, --pdb PDB — input PDB file name.

It is advised to edit the PDB file so that it contains only the atoms that will be used in the calculation (i.e. the CA atoms of each residue). The default selection used by dfs is “protein and name CA”, which selects all CA atoms of the protein(s).

ANM options

DFS uses an Anisotropic Network Model (ANM) to describe the protein dynamics. In ANM the protein is represented as a collection of nodes in the 3D space, corresponding to the residues’ alpha carbons. Two residues interact through a harmonic potential if the distance between them lies within a certain threshold. Therefore, the parameters that define the model are the distance cut-off that determines if two atoms are connected (r_c) and the force constant of the harmonic potential itself (γ).

-g, --gamma ANM_GAMMA — force constant value in kcal/mol/Å^2 (default: 1.0)

-c, --cutoff ANM_CUTOFF — distance cut-off in Å (default: 15)

-s, --selection SELECTION_STRING — atoms used in the ANM model (default: “protein and name CA”). ProDy selection strings can be used for atom selection (for more details see http://prody.csb.pitt.edu/manual/reference/atomic/select.html). This option can be used for example to select alpha carbons of specific chains in the PDB (e.g. “name CA and chain A”). Notice that ANM models are designed so that only alpha carbons should be considered.

-m, --write-hessian HESSIAN_OUTPUT — hessian matrix filename (optional output)
The ANM hessian matrix can be saved as a file in plain text format.

-M, --load-hessian HESSIAN_INPUT — hessian matrix filename (optional input)

A pre-calculated hessian matrix can be supplied by the user. If this option is used the hessian will not be calculated by dfs.

-i, --write-covariance COVARIANCE_OUTPUT covariance matrix filename (optional output) The covariance matrix calculated by inversion of the hessian can be saved as a text file.

-I, --load-covariance COVARIANCE_INPUT covariance matrix filename (optional input). A pre-calculated covariance matrix can be supplied by the user. If this option is used the covariance matrix will not be calculated by dfs.

Forces options

These options are used to define the number of DFS force orientations per site, the direction of the force vectors and, for Fixed Force (FF) calculations, their magnitude. Defaults settings are defined, but the user can supply a configuration file to define custom settings for both first and second sites:

-f CONFIG, --dfs CONFIG Force configuration file for DFS runs. Defaults will be used if such a file is not specified.

The format of the configuration files is explained in the template provided in the example directory. If this file is not supplied, default options will be used:

- number of forces: 12
- directions of forces: spherical Fibonacci lattice
- force magnitude: 10.0 kcal/mol/Å
- atoms to be considered as first sites: “protein and name CA”
- atoms to be considered as second sites: “protein and name CA”

Options for the rescuability score

The rescuability score can be calculated according to two different strategies. In the first mode (fixed_forces), all the forces are applied with the same magnitude (FF mode in the paper). In the second mode (fixed_displacements), force magnitudes are normalized so that they give rise to a fixed RMSD (~1 Å) between the original and the perturbed structure when a single force is applied (FR mode in the paper). By default, both modes are run in a dfs calculation.

-t [{fixed_forces,fixed_displacements,both}] Force mode to be used (default: both)

-d {drmsd,rmsd}, --conformational-distance {drmsd,rmsd} measure of conformational distance between the native and perturbed structures. It is possible to choose between root mean square deviation (rmsd) and distance root mean square deviation (drmsd) (default: drmsd)

-F {none,lsq_fit}, --fitting {none,lsq_fit} perform least square fitting between the perturbed and native structures. Useful for –d rmsd (default: none)

-a FIT_SELECTION, --fit-selection FIT_SELECTION selection string for the group used for fitting. Useful only in combination with –F lsq_fit (default: “protein and name CA”)

-x, --include-application-sites	Include force application sites in the calculation of the scores (by default they are not included)

Output

-S OUTPUT_SCORES, --output-scores OUTPUT_SCORES	Suffix for the filename of the rescuability score matrices. These are written in a user-readable ASCII number matrix format (default: “fixed_force_scores” or “fixed_displacements_scores”)

Additional information can be saved by the user in a binary file in HDF5 portable format (https://www.hdfgroup.org/). HDF5 files can be read with external software such as the HDF5 Tools (https://www.hdfgroup.org/products/hdf5_tools/) and hdfview, or selectively converted into a text file using the script data_muncher included in this dfs distribution. The content of the file can be changed with the option **-w**, which determines what will be actually written in the file:

-w [WRITABLE [WRITABLE ...]], --write [WRITABLE [WRITABLE ...]]	Space-separated list of data to be saved in the HDF5 details file. See below for a list of options, 'all' just saves everything. The flag -w will be ignored if -o is not supplied
-o OUTPUT_FNAME, --output OUTPUT_FNAME	Suffix to be used for the filename of the HDF5 details file. Only metadata are written if -w is not specified.
-q PRECISION, --precision PRECISION	Number of decimal places to be used in the HDF5 details file (-1 for lossless; default: 3)

By default, if the -o option is used without -w no information will be written in the HDF5 details file except from metadata for the run. The arguments of –w specify the type of additional information, stored using one matrix per residue pair. The file will contain the matrices of all the residue pairs involved in the calculations, labelled with the residue indices. The possible arguments of -w are:

- force_vectors: to save the force vectors as F*3N matrices, where F is the number of combinations of force orientations used for a given pair of residues and N is the number of residues. The components of the vector are 0 everywhere except for the positions corresponding to the residues where forces are applied.
- score_matrix: to save the N*N Rescuability score matrix

Others are more useful for developers and for having an in-depth look:

- perturbed_coordinates: to save the CA coordinates (in Å) of the perturbed structure (F*3N matrix).

- fitted_perturbed_coordinates: as in perturbed_coordinates, but after least square fitting has been performed. Only useful if fitting is actually performed (i.e. not useful with -d drmsd)
- displacements: to save the CA displacements Δ**R** induced by the force vectors (F*3N matrix)
- raw_score_matrix: to save the score matrix with the rescuability indices calculated for all the F combinations of force orientations.
- conformational_distance_matrices: to save the distance between native and perturbed structures calculated for all the F combinations of force orientations.
- scaling_factors: to save the scaling factors used for force magnitudes in the fixed_displacements mode
- all: to save all of the above

For instance, option "-w force_vectors perturbed_coordinates" will save the force vectors and the perturbed coordinates in the binary file. It should be noted that this file can get very big (in the order of tens or even hundreds of GB) and that writing to disk can thus become the bottleneck of the calculation, depending on your disk and processor speed.

2.2 compensatory_power

The compensatory_power script calculates the compensatory power using the FF and FR score matrices from dfs and the native structure of the protein. The compensatory power of a residue *j* is defined as $P^{FF/FR}(j) = N_S^{FF/FR}(j)/N_C(j)$, where $N_S^{FF/FR}(j)$, is the number of residues for which $S_{ij}^{FF/FR} > 0$ and $N_C(j)$ is the number of contacts of j (i.e. the number of CA atoms within d of j). The compensatory power can be calculated singularly for each type of score matrix (either FF or FR) or an overall compensatory power can be calculated by providing both matrices in input.
When option -P is used, the compensatory power values are written in the B-factor field of the PDB file (values are first multiplied by 10 to avoid loss of precision).

The script requires:

-p PDB, --pdb PDB — PDB filename. The same file provided for the dfs run should be provided.

-m SCORE_MATRIX [SCORE_MATRIX ...] — Score matrix filename(s)

-s SEL_STRING [SEL_STRING ...] — atom selection (default: "protein and name CA"). The same selection used for the dfs run (-s option) should be provided.

Other options include:

-c SCORE_CUTOFF, --score-cutoff SCORE_CUTOFF — Rescuability score cut-off to detect rescued positions (default: 0.0)

-d DISTANCE_CUTOFF — Distance cut-off to identify contacts (Å; default: 15.0)

Output options include:

-o OUTFILE, --output OUTFILE — Name for the output file (default: normalized_score.dat) containing the compensatory scores for the residues selected in the calculation

-P PDB_OUTFILE, --pdb-output PDB_OUTFILE — Name for the output PDB file with the compensatory power values saved in the B-

factor column (default: do not save)

2.3 dotM

dotM calculates the Root Mean Square Inner Product (RMSIP) between the ANM normal modes $\mathbf{n}_l$ and a displacement vector $\Delta\mathbf{R}$ calculated by dfs as the the difference between the structure perturbed by two forces and either the unperturbed structure (-m option) or the structure perturbed by a single force (default). RMSIP for this specific case is calculated as:

$$RMSIP = \sqrt{\sum_{l=1}^{m} (\Delta \boldsymbol{R} \cdot \mathbf{n}_l)^2}$$

RMSIP is a measure of overlap between the two different sets of vectors and ranges from 0 to 1, where 1 indicates perfectly superimposable sets. The dotM script also calculates the Shannon entropy from the distribution of the normal modes that have best overlap with the displacements involving a given second site. The script requires:

-f H5_FNAME, --details H5_FNAME HDF5 details file from a DFS calculation. This file must have been generated in a dfs run with -w raw_score_matrix and either perturbed_coordinates or fitted_perturbed_coordinates

-p ORIGINAL_PDB, --pdb ORIGINAL_PDB Original unperturbed PDB file

-R [FIRST_SITE [FIRST_SITE ...]], --first-sites [FIRST_SITE [FIRST_SITE ...]] First sites to be considered, see below

-D [SECOND_SITE [SECOND_SITE ...]], --second-sites [SECOND_SITE [SECOND_SITE ...]] Second sites to be considered, see below

Options -R and -D accept input in multiple ways. Residues can be specified a) by the residue serial number, with residues numbered sequentially starting from 0 according to their order in the PDB file, b) by a string containing the PDB chain and residue sequence number (for instance, A34 for the residue with resSeq 34 and chainID A in the PDB file), c) for selection of multiple residues, by supplying a selection string in ProDy style. For multiple residue selections, all the possible combinations of -R and -D residues are considered. When the -R and -D selections overlap, the calculation for “self” (i, i) pairs is not performed and “NA” (not available) is printed instead of the RMSIP value.

The direction of the forces producing the displacements vector $\Delta\mathbf{R}$ can be specified with:

-r FS_DIRECTION, --first-site-direction R_FS_DIRECTION Direction of the force on the first site ('max' or number; default: 'max'). This option accepts a number from 0 to N-1, where N is the number of force directions stored in the HDF5 file. The keyword "max" selects the direction that produces the maximum rescuability index.

-d SS_DIRECTION, --second-site-direction SS_DIRECTION Direction of the force on the second site ('max' or number; default: 'max'). Works like option –r.

Other options include:

-o OUTPUT_FILE, --output-rmsip OUTPUT_FILE name of text output file for RMSIP values (default: rmsip.txt)

-e ENTROPY_FILE, --output-entropy ENTROPY_FILE name of text output file for Shannon entropies (default: entropy.txt)

-m, --pdb-ref-coords Use the coordinates in the PDB instead of the single-force structures to calculate displacements

-n N_MODES number of ANM modes to be used for the calculation of RMSIP (default: 10)

-t {perturbed_coordinates,fitted_perturbed_coordinates}, --data-type {perturbed_coordinates,fitted_perturbed_coordinates} data type to be used for the calculation of displacements (default: perturbed_coordinates)

2.4 data_muncher

data_muncher extracts data from the HDF5 details file (binary format) and saves them as numerical data files in .npz compressed or text format. Protein structures are saved as .dcd trajectory files. The script requires:

-f HDF5, --details HDF5 details file produced by dfs

-p PDB, --pdb PDB Original PDB structure

-R FIRST_SITE [FIRST_SITE ...], --first-sites FIRST_SITE [FIRST_SITE ...] First site (default: protein and name CA)

-D SECOND_SITE [SECOND_SITE ...], --second-sites SECOND_SITE [SECOND_SITE ...] Second site (default: 'name CA')

-w [WRITABLE [WRITABLE ...]] Choose one or more data type that should be saved as the output files. If no option is given, data_muncher will just display metadata and quit.

The option **-w** works similarly as the same option of dfs, but in this case it allows the user to select what should be extracted from the details file. All the keywords accepted for -w in dfs work here and have the same meaning. Additionally, -w in data_muncher accepts:

- max_score_perturbed_coordinates: writes two PDB files containing 1) the perturbed structure generated by the combination of forces with the maximum rescuability index and 2) the corresponding single-force perturbed structure
- max_score_fitted_perturbed_coordinates: same as max_score_perturbed_coordinates, but using the perturbed coordinates fitted to the native structure
- all_available: all the available data in the details file

Data can be extracted by data_muncher only if they are present in the HDF5 compressed file, so -w arguments of data_muncher should always be a subset of arguments of -w in dfs, except that for the additional keywords.

Other options control the output format:

-t, --textual Write data in simple text format instead of npz

-d DECIMAL, --decimal DECIMAL Number of decimal places for numeric data (default: 3)