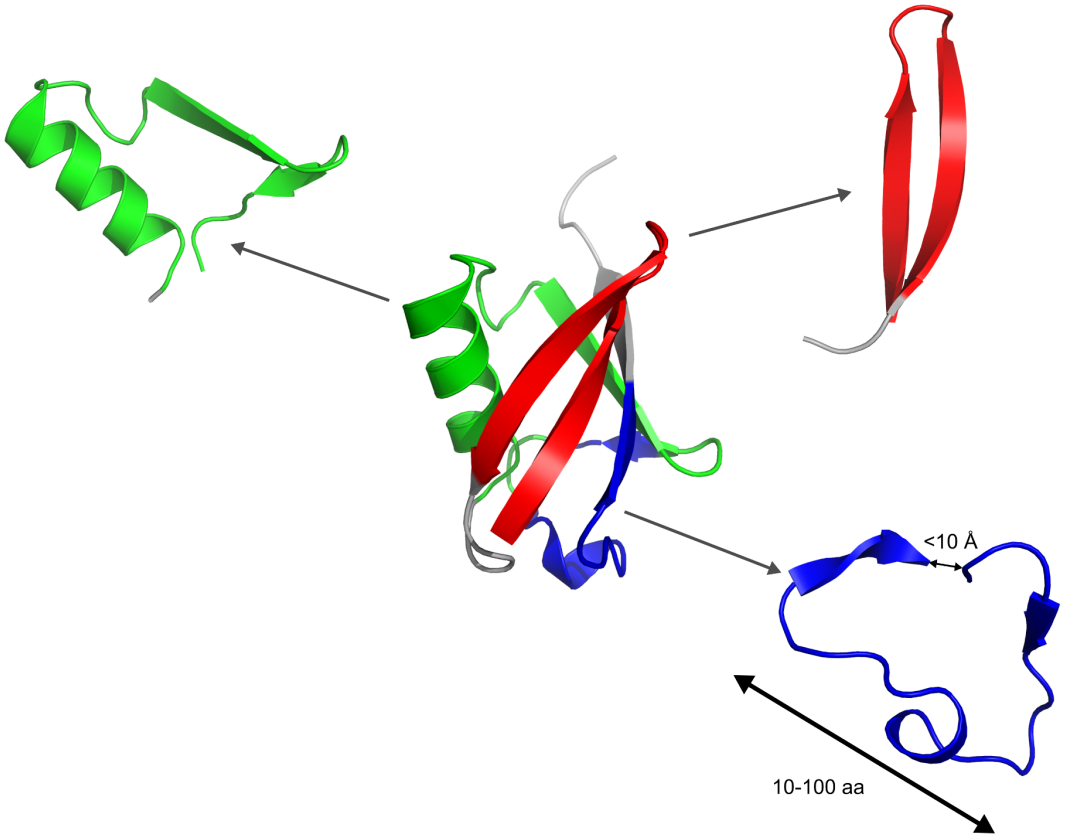
TEF 2.0 – User manual

2018-06-27

D. Stratmann, JS. Pathmanathan, G. Postic, J. Rey, J. Chomilier



Sorbonne Université, UMR 7590 CNRS, MNHN, IRD, Institut de Minéralogie de Physique des Matériaux et de Cosmochimie (IMPMC), Paris, France

INSERM UMR-S 973, Université Paris Diderot, Sorbonne Paris Cité, RPBS, Paris, France

Contact: dirk.stratmann@sorbonne-universite.fr

Contents

1 INTRODUCTION **3**

2 TEF 1.0 versus TEF 2.0 **4**

3 How to use TEF 2.0 ? **4**

4 OPTIONS [*name of the option (command line)*] **5**

4.1 PDB files . . . 5

4.1.1 One single PDB file (-pdbFile filename) . . . 5

4.1.2 Directory with PDB files (-pdb directory) . . . 5

4.2 Output files . . . 5

4.2.1 Output directory (-out directory) . . . 5

4.2.2 Sub-directories (-subdir) . . . 5

4.2.3 Run pymol (-pymol) . . . 5

4.2.4 Debug information (-debug) . . . 5

4.3 TEF selection approach (-version 2) . . . 5

4.4 Common parameters for both selection approaches (TEF 1.0 and TEF 2.0) . . . 6

4.4.1 Maximum TEF ends distance (-d 10.0) . . . 6

4.4.2 Minimum TEF length (-min 10) . . . 6

4.4.3 Maximum TEF length (-max 100) . . . 6

4.5 TEF 2.0 parameters . . . 6

4.5.1 Scores weights . . . 6

4.5.2 Overlap (-overlap 2) . . . 6

4.5.3 MAX_GAP (-gap 100) . . . 7

4.5.4 Use NACCESS (-naccess) . . . 7

4.5.5 Residues accessibility (-a 25) . . . 7

5 RESULTS **7**

5.1 TEF representation [XXXX_solutions.tef, XXXX = PDB id] . . . 7

5.2 List of all possible TEFs [XXXX_positions_chain_X.tef, XXXX = PDB id and X= chain name] . . . 8

5.3 Pymol script + PNG image [XXXX.pymol/png, XXXX = PDB id] . . . 9

5.4 Input parameters [parameters.tef] . . . 9

5.5 Statistics [all_xxxxxxx.tef] . . . 9

5.5.1 Solutions [all_results.tef] . . . 9

5.5.2 Sequence coverage [all_coverage.tef] . . . 9

5.5.3 Average TEF ends distances [all_ca_dist_mean.tef] . . . 9

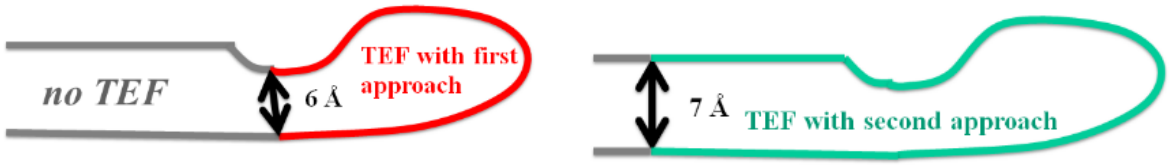
5.5.4 TEF lengths [all_tef_length.tef] . . . 10

5.5.5 Number of TEFs [all_mean_tef.tef] . . . 10

1 INTRODUCTION

A globular protein is composed of long chains of amino acids folded occasionally on itself, forming loop like trajectories with typical Cα-Cα distances below 10 Å between the ends (fig.1). These fragments were initially called closed loops [1]. The histogram of the sequence separation between these contact residues presents a maximum at around 20 - 30 amino acids [2]. Later on, it was shown that the ends of these closed loops include mainly hydrophobic residues [3], and a thorough analysis demonstrated that these hydrophobic amino acids were highly conserved among structures of the same family even among distantly related members, and they were called topohydrophobic positions [4]. The concept of TEFs (Tightened End Fragments) emerged from the joint concepts of closed loops and topohydrophobic positions [5].

The decomposition of a protein structure into TEFs can be done in different ways because of redundancy due to overlapping TEF. This redundancy is quite huge, as the length of all possible TEF exceeds about 50 times the sequence length. The first approach *(distance approach, TEF 1.0 program)* selects the TEFs with the tightest ends in terms of distance between Cα-atoms (fig.2a).This approach is also used by the DHcL server from Berezovsky et al. [6]. The disadvantage of this approach is that the poor coverage of the protein by TEFs. In order to improve the splitting of a domain into its constituting TEFs, the second approach *(score approach, TEF 2.0 program)* selects a sequence decomposition by TEFs that minimizes the number of residues unassigned to any TEF and choose TEFs with the tightest ends in terms of distance between Cα-atoms (fig.2b).



(a) (b)

Figure 2: TEF with (a) distance approach (TEF 1.0 program) and (b) score approach (TEF 2.0 program)

For this task a graph-algorithm tests all combinations of TEFs yielding the optimal decomposition according to a score. The best solution is the solution with the lowest score. The score is composed of three individual scores, $Score_{gap}$ which measure the percentage of residues unassigned to any TEF (gap), $Score_{c\alpha-c\alpha}$ which measure the mean of the distances between Cα-atoms and a third optional $Score_{frag}$ which measure the number of TEFs (fragmentation) in the solution.

- The gap score $Score_{gap}$ is simply the sum of gaps, i.e. the number of residues unassigned to any TEF :

$$Score_{gap} = \sum gap * w_{gap} \tag{1}$$

where w_{gap} is the weight (default : 1.0) for this score.

- The distance cα-cα score $Score_{c\alpha-c\alpha}$ is the sum of differences to an average distance d_{avg} :

$$Score_{c\alpha-c\alpha} = \sum_{i=0}^{N_{TEF}} (d_{c\alpha-c\alpha} - d_{avg}) * w_{c\alpha-c\alpha} \tag{2}$$

where $d_{c\alpha-c\alpha}$is the distance between Cα-atoms of a TEF i, N_{TEF} is the total number of TEFs in the solution and $w_{c\alpha-c\alpha}$is the weight (default : 1.0) for this score. d_{avg} is simply the middle of the interval

of distances $\left[\frac{d_{max}}{2}, d_{max}\right]$, with d_{max} being the maximum allowed distance.

- The optional fragmentation score $Score_{frag}$ can be included to obtain a higher or lower fragmentation, i.e. smaller or longer TEFs on average. Its formula is :

$$Score_{frag} = -N_{TEF} * 10 * w_{frag} \tag{3}$$

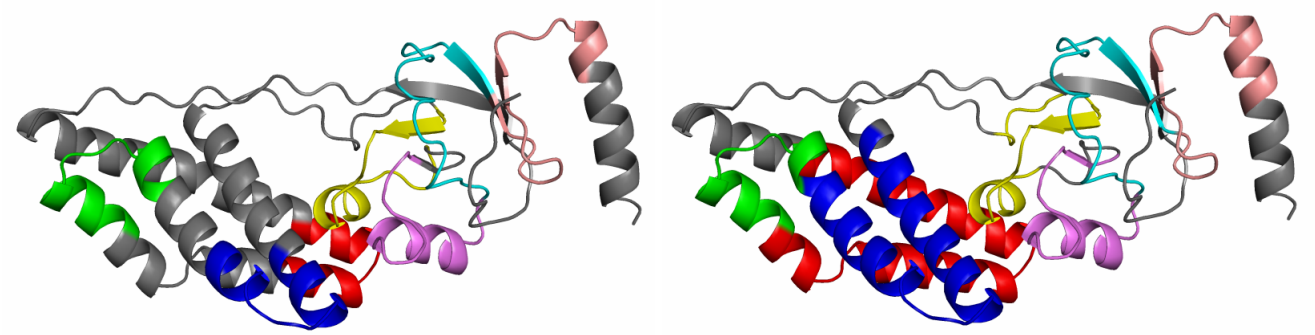
where w_{frag} is the weight (default : 0.0) for this score. For $w_{frag} > 0$ a higher fragmentation will be obtained and for $w_{frag} < 0$ a lower fragmentation will be obtained

The final score for a solution is the sum of the three scores :

$$Score_{solution} = Score_{gap} + Score_{c\alpha-c\alpha} + Score_{frag} \tag{4}$$

2 TEF 1.0 versus TEF 2.0

We have tested the two approaches - TEF 1.0 and TEF 2.0 - on a data base composed of 278 proteins with less than 25 % sequence identity in order to compare their coverage rate by TEFs on proteins. The distance approach (TEF 1.0) gives an average coverage rate of only $(67 \pm 8)\%$ while the score approach (TEF 2.0) attains $(95 \pm 3)\%$ (fig.3). At the same time the average distance between the TEF ends increased only from $(5.1 \pm 0.3)\mathring{A}$ (TEF 1.0) to $(6.4 \pm 0.6)\mathring{A}$ (TEF 2.0). If required, the TEF ends distance based optimization done by of TEF 1.0 can be done also by TEF 2.0 by setting the weight w_{gap} for the gap score equal to zero (command line: -gw 0).



(a) (b)

Figure 3: Decomposition into TEFs of N terminal parts of Enzyme I (PDB code : 3EZA) with (a) distance approach (TEF 1.0) and (b) score approach (TEF 2.0)

3 How to use TEF 2.0 ?

Step 1 : Give PDB file or enter the PDB code of the protein.

Step 2 : Choose an approach (TEF 1.0 or TEF 2.0) of the decomposition of protein into TEFs (see section OPTIONS).

Step 3 : Depending on your needs you can change the default values of options for the selected approach (see section OPTIONS) or keep them.

Step 4 : click on run and you will obtain the results.

4 OPTIONS [*name of the option (command line)*]

4.1 PDB files

Without any option, all PDB files in the current directory will be treated by the TEF program.

4.1.1 One single PDB file (-pdbFile filename)

If only a specific PDB file should be treated by the TEF program, use the -pdbFile option to specify the path and filename of the PDB file. This PDB file can also be an assembly of several PDB files put into one single "multi-PDB" file.

4.1.2 Directory with PDB files (-pdb directory)

If the PDB files are not in the current directory, specify with the -pdb option the path to the directory containing the PDB files to be treated by the TEF program.

4.2 Output files

4.2.1 Output directory (-out directory)

The output files are stored by default in the current directory. This can be changed by specifying the output directory with the -out option.

4.2.2 Sub-directories (-subdir)

The output files can be stored automatically in different sub-directories, one per PDB file (default: no sub-directories).

4.2.3 Run pymol (-pymol)

With this option TEF 2.0 will call pymol to generate a PNG file for each PDB. The TEFs are indicated by colors (see section 5.3). As the ray-tracing step can take a bit of time, the generation of the PNG file is deactivated by default.

4.2.4 Debug information (-debug)

Additional files for debugging purposes will be written.

4.3 TEF selection approach (-version 2)

The distance approach (-version 1) selects the TEFs with the tightest ends in terms of distance between Cα-atoms (fig.2a).
The score approach (-version 2, default) selects a sequence decomposition by TEFs that minimizes the number of residues unassigned to any TEF and choose TEFs with the tightest ends in terms of distance between Cα-atoms (fig.2b).

4.4 Common parameters for both selection approaches (TEF 1.0 and TEF 2.0)

4.4.1 Maximum TEF ends distance (-d 10.0)

Maximum distance between the Cα-atoms of the the TEF ends.
Default : 10.0 Å
Allowed : 4.0 - 15.0 Å

4.4.2 Minimum TEF length (-min 10)

The minimum length of a TEF can be specified to avoid too short fragments.
Default : 10 AA
Allowed : 10 - 100 AA

4.4.3 Maximum TEF length (-max 100)

The maximum length of a TEF can be specified to avoid too long fragments.
Default : 100 AA
Allowed : 10 - 100 AA

4.5 TEF 2.0 parameters

4.5.1 Scores weights

It's possible to change the weight of each scores (formulas 1, 2 and 3) and modify the decomposition of the protein structure into TEFs.

- **Degree of Coverage, w_{gap} (-gw 1.0)**

The solution will favor a higher coverage by TEFs for higher values.
Default: 1.0
Allowed : 0.0 – 100.0

- **Cα-atoms distance weight, $w_{c\alpha-c\alpha}$ (-dw 1.0)**

The solution will favor TEFs with shorter distances at their end for higher values.
Default: 1.0
Allowed : 0.0 – 100.0

- **Degree of fragmentation, w_{frag} (-tw 0.0)**

The solution will favor a higher/lower number of short TEFs for positive/negative values.
Default : 0.0
Allowed : -100.0 – 100.0

4.5.2 Overlap (-overlap 2)

By default the maximum allowed overlap between two TEFs is 2 residues, which can be changed by this option.

4.5.3 MAX_GAP (-gap 100)

The MAX_GAP value controls the search depth for the optimal selection of TEFs. MAX_GAP corresponds to the maximal length between two TEFs. The gaps are counted from the first possible TEF after the last residue of the current TEF, in order to jump over large part of the sequence without any TEF. A minimum value of 10 is recommended for MAX_GAP, a too small value may result in an incomplete search. Higher values will result in a longer search time, but not necessarily change the final result.

Default : 100 AA

Allowed : 0 - 300 AA

4.5.4 Use NACCESS (-naccess)

With this option the NACCESS [7] program can be used to restrict the TEF-ends to the protein core.

By default this filter for possible TEFs is deactivated.

4.5.5 Residues accessibility (-a 25)

The relative maximum Accessible Surface Area (ASA) of the TEF-ends is calculated by NACCESS [7]. A small value (< 50%) will constrain the TEF-ends to the protein core, if the option **-naccess** is also used.

Default : 25 %

Allowed : 1 - 200 % (for some cases NACCESS gives values > 100%)

5 RESULTS

5.1 TEF representation [XXXX_solutions.tef, XXXX = PDB id]

In the figure 4 is represented the sequence of the submitted structure and just below the corresponding TEFs. There are two lines of TEFs to better visualize when two TEFs overlap (maximum overlap 2 residues). Below are listed the TEFs (first residue, last residue and distance in Å).

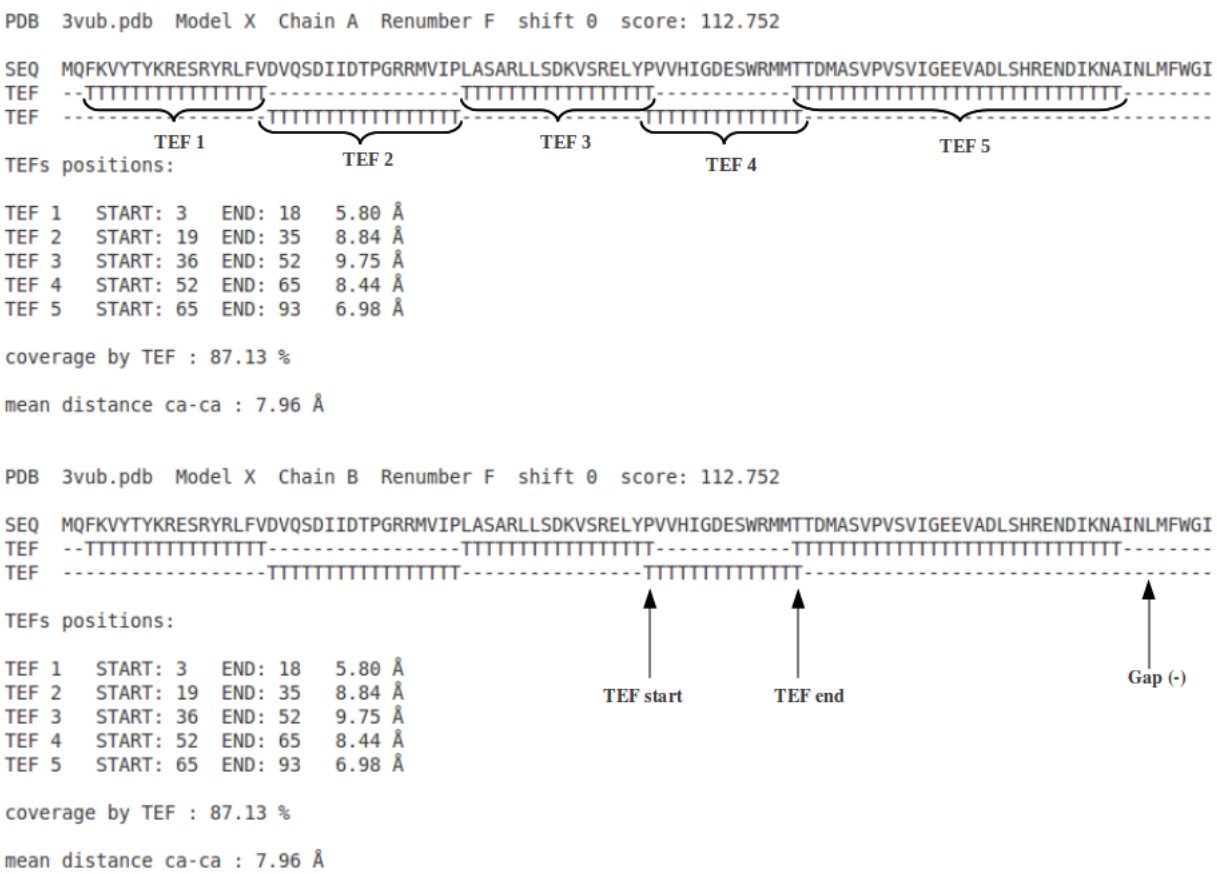


Figure 4: Output of TEF program for TEF representation

5.2 List of all possible TEFs [XXXX_positions_chain_X.tef, XXXX = PDB id and X= chain name]

The list of all possible TEFs is represented (fig.5) like this :

- column 1 : first residue of the TEF
- column 2 : last residue of the TEF
- column 3 : size (in residues) of the TEF
- column 4 : distance in Å between the first and last residues of the TEF

```
start  end  length  distance ca-ca

3      17   14      9.48 Å
3      18   15      5.80 Å
3      19   16      5.30 Å
3      21   18      8.20 Å
3      33   30      9.16 Å
3      34   31      9.89 Å
...    ...  ...     ...
54     90   36      9.94 Å
56     83   27      9.11 Å
63     90   27      8.36 Å
63     93   30      9.07 Å
64     93   29      9.85 Å
65     90   25      8.68 Å
65     93   28      6.98 Å

Mean length: 29
```

Figure 5: Output of TEF program for all possible TEFs list

5.3 Pymol script + PNG image [XXXX.pymol/png, XXXX = PDB id]

A pymol [8] script is generated which allowed to visualize in 3D the decomposition of the protein into TEFs. For that you have to download the cleaned PDB and the script then execute the pymol script. This script will write at the end a PNG image file (XXXX.png, XXXX = PDB id). With the **-pymol** option the TEF program will generate automatically this PNG image file along with the pymol script. The Figure 6 shows the color code corresponding to the TEF ids.

TEF NUMBER	Pymol color
TEF 1	red
TEF 2	green
TEF 3	blue
TEF 4	yellow
TEF 5	violet
TEF 6	cyan
TEF 7	salmon
TEF 8	lime
TEF 9	pink
TEF 10	slate
TEF 11	magenta
TEF 12	orange
TEF 13	marine
TEF 14	olive
TEF 15	purple
TEF 16	teal
TEF 17	forest
TEF 18	firebrick
TEF 19	chocolate
TEF 20	wheat
TEF 21	red
TEF 22	green
TEF 23	...
... ...	...

Figure 6: TEF colors in pymol

5.4 Input parameters [parameters.tef]

The parameters.tef file contains the command line as well as a list of all parameters used in the run.

5.5 Statistics [all_xxxxxxx.tef]

The files beginning with "all_" summarize the results for all PDB files/chains that has been treated by the TEF program in the current run.

5.5.1 Solutions [all_results.tef]

Gives the TEF decomposition solutions of all PDB files/chains in a compact form.

5.5.2 Sequence coverage [all_coverage.tef]

The list of sequence coverage values (in %) for all PDB files.

5.5.3 Average TEF ends distances [all_ca_dist_mean.tef]

The list of average TEF ends distance values (in Å) for all PDB files.

5.5.4 TEF lengths [all_tef_length.tef]

A list of TEF lengths (in number of residues) of all selected TEFs for all PDB files.

5.5.5 Number of TEFs [all_mean_tef.tef]

Gives a list for all PDB files of the number of selected TEFs.

References

[1] Varda Ittah and Elisha Haas. Nonlocal interactions stabilize long range loops in the initial folding intermediates of reduced bovine pancreatic trypsin inhibitor. *Biochemistry*, 34(13):4493–4506, April 1995.

[2] I N Berezovsky, A Y Grosberg, and E N Trifonov. Closed loops of nearly standard size: common basic element of protein structure. *FEBS Letters*, 466(2-3):283–286, January 2000.

[3] I N Berezovsky, V M Kirzhner, A Kirzhner, and E N Trifonov. Protein folding: looping from hydrophobic nuclei. *Proteins*, 45(4):346–350, December 2001.

[4] Anne Poupon and Jean–Paul Mornon. Populations of hydrophobic amino acids within protein globular domains: Identification of conserved 'topohydrophobic' positions. *Proteins: Structure, Function, and Bioinformatics*, 33(3):329–342, November 1998.

[5] M Lamarine, J P Mornon, N Berezovsky, and J Chomilier. Distribution of tightened end fragments of globular proteins statistically matches that of topohydrophobic positions: towards an efficient punctuation of protein folding? *Cellular and Molecular Life Sciences: CMLS*, 58(3):492–498, March 2001.

[6] Grzegorz Koczyk and Igor N Berezovsky. Domain hierarchy and closed loops (DHcL): a server for exploring hierarchy of protein domain structure. *Nucleic Acids Research*, 36(Web Server issue):W239–245, July 2008.

[7] Hubbard, S.J. and thornton, J.M. (1993), "NACCESS", computer program, department of biochemistry and molecular biology, university college london.

[8] Delano, W.L. The PyMOL Molecular Graphics System (2002) DeLano Scientific, san carlos, CA, USA. http://www.pymol.org.