*D374–D379* *Nucleic Acids Research, 2009, Vol. 37, Database issue* *Published online 22 October 2008*
*doi:10.1093/nar/gkn704*

# SPROUTS: a database for the evaluation of protein stability upon point mutation

**Mathieu Lonquety[1,2], Zoé Lacroix[1,2,3,*], Nikolaos Papandreou[4] and Jacques Chomilier[2]**

[1]Scientific Data Management Laboratory, Arizona State University, Tempe AZ 85282-5706, USA, [2]Protein Structure Prediction, IMPMC, Université Paris 6, CNRS UMR 7590, 140 rue de Lourmel, 75015 Paris, France, [3]Pharmaceutical Genomics Division, Translational Genomics Research Institute, 13400 E Shea Blvd, Scottsdale AZ 85259, USA and [4]Laboratory of Genetics, Agricultural University of Athens, Iera Odos 75, 118 55 Athens, Greece

Received August 1, 2008; Revised September 26, 2008; Accepted September 29, 2008

## ABSTRACT

**SPROUTS (Structural Prediction for pRotein fOlding UTility System) is a new database that provides access to various structural data sets and integrated functionalities not yet available to the community. The originality of the SPROUTS database is the ability to gain access to a variety of structural analyses at one place and with a strong interaction between them. SPROUTS currently combines data pertaining to 429 structures that capture representative folds and results related to the prediction of critical residues expected to belong to the folding nucleus: the MIR (Most Interacting Residues), the description of the structures in terms of modular fragments: the TEF (Tightened End Fragments), and the calculation at each position of the free energy change gradient upon mutation by one of the 19 amino acids. All database results can be displayed and downloaded in textual files and Excel spreadsheets and visualized on the protein structure. SPROUTS is a unique resource to access as well as visualize state-of-the-art characteristics of protein folding and analyse the effect of point mutations on protein structure. It is available at http://bioinformatics.eas.asu.edu/sprouts.html.**

## INTRODUCTION

The production of point mutation in a sequence is now routinely performed in molecular biology laboratories since the development of protein-engineering techniques. In the field of fundamental research, it is widely used in order to verify whether a given amino acid belongs to the folding nucleus supported by the $\Phi_F$ value determination initially proposed by Fersht (1,2). Indeed mutations may have unexpected yet significant impact. For example, an overexpression of eukaryotic sequences in *Escherichia coli* may produce inclusion bodies instead of soluble globules. One way to avoid this problem is to create random mutations, hoping that the solubility will be increased. However, one has to check whether the proposed mutations have dramatic effects such as greater instability which may lead in some cases to an unfolded protein or to inclusion bodies.

If protein stability changes upon point mutation have given rise to the development of prediction programs such as the ones used in this work, few data have been collected and proposed to the scientific community. One can cite the Protherm database (3) which contains thermodynamic experimental data including free energy changes or the Protein Mutant Database (4) which includes references to mutant proteins from the literature, but no database devoted to the collection of stability changes prediction exists, to the best of our knowledge. There exists few databases devoted to protein folding, a field in great expansion [see for example Protein Folding Database (5)] but none provides free energy calculation or the two original concepts we propose, MIR (Most Interacting Residues) and TEF (Tightened End Fragments). Therefore, a database containing predictions of stability, and their evaluation, supports critical applications of both fundamental and experimental research.

## MATERIALS AND METHODS

### MIR prediction

MIR prediction is achieved as follows (6). An algorithm devoted to the simulation of the early steps of protein folding has been developed. It is based on a (2,1,0) lattice coupled with a Monte Carlo algorithm. Amino acids are distributed at random on the nodes of the lattice from which they can move to an unoccupied node. Initial and final conformation energies are calculated with the Miyazawa and Jernigan potential of mean force (7) and

*To whom correspondence should be addressed. Tel: +1 480 727 6935; Fax: +1 480 965 8325; Email: zoe.lacroix@asu.edu

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

the Metropolis criterion is applied either to validate the move or not. The simulation is stopped in the early stages of the process, typically $10^6$ Monte Carlo steps, and this process is repeated 100 times with different random initial conformations. The Number of Contact Neighbours (NCN) in the lattice for each residue is periodically recorded and averaged over the simulation times. It produces a mean number of first neighbours for each amino acid of the protein structure reflecting the level of interaction involved. All amino acids above a given threshold of first neighbours are called MIR. Hydrophobic at more than 90%, MIR statistically correspond to the residues constituting the cores of the proteins (6). The algorithm can be run on the RPBS server (8). More details are available on the help section of the website: http://bioinformatics.eas.asu.edu/springs/Sprouts/projectsSproutsFAQ.html#faq_MIR.

### TEF assignment

Structures are all analysed under the principle of a succession of fragments with their ends close in the 3D space with a typical distance between their alpha carbons below 10 Å. The idea is related to the paradigm of the autonomous folding units (9). Indeed, Berezovsky *et al.* (10) have demonstrated that structures can be split in successive fragments of mean length of 25 amino acids. These fragments have been previously described as closed loops (11). Moreover other studies have concluded that the extremities of the previous closed loops were occupied by hydrophobic residues highly conserved among members of a functional family at the so-called topohydrophobic positions (12). The assumption is that these ends are located at the core of the globular protein and we have already shown that they are close to MIR positions (6). Finally, the conjunction of closed loops and topohydrophobic positions has given rise to the TEF (13). See the help section of the website: http://bioinformatics.eas.asu.edu/springs/Sprouts/projectsSproutsFAQ.html#faq_TEF for more details.

### Stability calculation

Stability is evaluated by five programs publicly available: DFIRE (14), two versions of I-Mutant (15), MUpro (16) and PoPMuSiC (17). Other methods exist but have been rejected due to some restrictions: CUPSAT (18) was not available in a stand-alone version and the current version of FoldX (19) only computes mutations to Alanine. Eris (20) and AUTO-MUTE (21) were not published at the time we began our project (February 2007). The evaluation of the five tools is accomplished by reference to Protherm (3) the only database providing experimental data. It collects information from the literature over perturbation of stability due to point mutation, on the basis of free energy change. A score is produced in order to homogenize the various algorithms from –19 for very stable positions (in other words, a mutation is unexpected) to the theoretical upper limit of + 19, corresponding to very unstable position. This latter case has not been evidenced, a consequence of the optimization of wild-type sequences due to evolution. Finally, we also developed a consensus which is a mean of the stability scores given by the five tools. In the case where one single piece of information is missing, this consensus is not calculated.

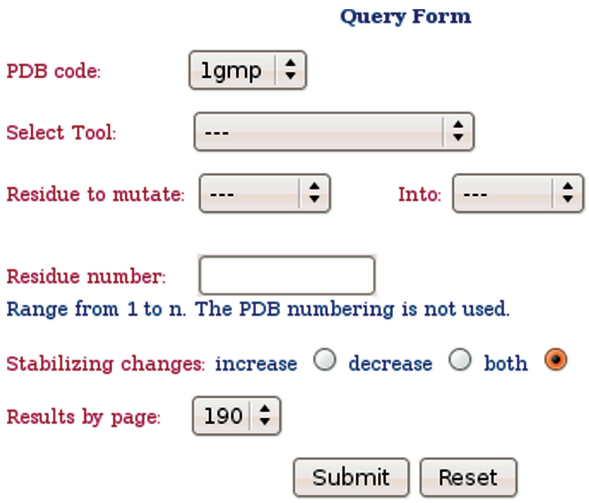


**Figure 1.** Query interface of SPROUTS with 1gmp as selected structure and default options.

### Construction of the database

SPROUTS is designed to support various studies that involve stability computations. Protein structure analysis is computationally expensive, as the execution time is typically of 5 hours for one sequence. Besides, on most of the sites of the algorithms that are used in this study, one query corresponds to one single mutation, therefore the user needs to fill 100 requests for a sequence of 100 residues (considering that the tool processes all the 19 possible mutations which is not the case for all of them).

The database is organized in three tables: Tools, Proteins and Results which gather information on 429 proteins (77 124 amino acids) with prediction of free energy difference changes for each residue and each possible mutation computed for the five tools previously described.

### Querying the database

Inquiries are done with a PDB code (22). The algorithm name, the position to mutate and the mutation can be selected at will. The default parameters return the free energy change for all substitutions, on all positions, for the five algorithms. One can decide to reduce the query to one given mutation at one single position for one of the tools. Figure 1 represents the query interface with the 1gmp structure selected and default options.

In addition, one can select the number of results per page and also results regarding their interpretation of $\Delta\Delta G$, i.e. whether the mutation is more favourable in terms of energy than the wild-type residue or not.

### Visualization

The output can be visualized in tables which collect all the raw results, i.e. $\Delta\Delta G$, and that can be downloaded by the user. Two visualization modes enhance the access to the results: a 2D mode displays graphs that summarize

the stability score whereas a 3D mode represents the results directly on the 3D protein structure.

*2D mode*. The stability graphs summarize the stability score for a whole sequence and for designated tools. Because the stability score curves are very sharp and hardly interpretable, a smoothing option has been set by default. Based on the Pascal triangle method, this technique takes into account the neighbourhood of a point (four neighbours from each side of the point in this case) reducing the number of peaks. The counterpart is the loss of accuracy but it helps to localize the regions of interest. The upper window in Figure 2 shows the type of graph one can obtain on the server and the information related to the TEF assignment and MIR prediction. The stability graph is useful to visualize and quickly localize the sequence regions very sensitive to mutations. Indeed, if almost every mutation has a destabilizing effect, the score will be close to –19. Conversely, if almost any mutation has a stabilizing effect, the score will be positive. The ability to distinguish these extremes is of great importance as it highlights the positions that should play a role in the folding of the structure.

*3D mode*. The third visualization mode benefits structural biologists who are more accustomed to manipulating 3D structures and objects. This feature offers the possibility to display the structure in 3D and to emphasize the three sources of information contained on the server. First, one can represent the stability score of each alpha carbon of the protein with a colour gradient in the range red for –19 up to blue for +19. Secondly, MIR can be represented as partially transparent purple spheres around designated alpha carbons. Finally, the best way to represent TEF is colourizing the cartoon representation of the structure. Each colour is associated to a different TEF and when an overlap occurs, the intermediate colour between the two ones involved is drawn. The present representation may give too much information to be easily interpreted and we are currently working on a more sophisticated query process to capture the minimal data required by a user to answer his question. The lower window in Figure 2 shows an example of the 3D applet with the 1gmp structure.

## Use case

We illustrate the use of the database with a use case. A typical question SPROUTS can answer regards the feasibility of designing a stable mutant for a given protein without weakening its structure. In particular, it can inform on the positions critical to the maintenance of the structure and that should not be mutated. The example of a ribonuclease from *Streptomyces aureofaciens* (PDB code: 1gmp) illustrates this case. We present here a summary of this use case and a fully detailed version is available as Supplementary Data and available at http://bioinformatics.eas.asu.edu/sprouts-case.html.

The query process consists in selecting the 1gmp structure in the PDB code field and using the default options as illustrated in Figure 1. As we are looking for highly destabilizing mutations on any position along the sequence, parsing the raw results is a tedious task. Indeed, it results in 48 pages combining 9101 different $\Delta\Delta G$ which can however be downloaded in 'csv' format for further analysis with any spreadsheet manager software.

The development of different visualizing modes has been initiated to reduce the time needed to retrieve, parse and analyse the results and answer the type of question we are formulating here. By clicking on the 2D mode button in the main result page, a pop-up appears and the user can locate the positions of interest along the whole sequence. Figure 2 shows the pop-up window with the results obtained for the DFIRE and I-Mutant sequence + structure tools on the queried protein. We focus on two tools in order to simplify the description of the querying and analysis scenario. The smoothing process has been applied at the exception of the N- and C-terminal ends, because of the window size on which data are smoothed. The user is looking for positions corresponding to stability score minima and these characteristics are emphasized in the graph zone. Note that the sequence position numbering displayed goes from 1 to 96. One can count 11 minima for DFIRE on the following positions: 7, 20, 28, 37, 44, 52, 56, 71, 81, 86, and 92. I-Mutant sequence + structure have 12 minima on the following positions: 7, 11, 22, 27, 35, 44, 52, 56, 60, 71, 82, and 93.

If one now looks at the MIR prediction located under the graph zone on the MIR line with the character M as residues predicted as MIR, residues 8, 22, 36, 57, 58, 70, 71, 86, 89, 91, 92, and 96 are concerned. In the worst case, if one does not include any flexibility by retrieving only the exact matches, the sole residue 71 is considered as a minimum of stability for both tools and characterized as a MIR. Moreover, if one looks at the TEF assignment (TEF fragments are represented on two lines TEF with strings of T), this position corresponds to a TEF end. Indeed, it confirms the structural importance of this amino acid as it has been demonstrated that TEF ends are located in the core of the protein and thus play a role in the maintenance of its conformation (13). When computing the solvent-accessible surface, it appears that Ile71 is completely buried, with a relative accessible area of 0%. Finally, in the 3D mode pop-up, this position corresponds to one of the purple spheres located in the core of the protein on the middle strand of the β sheet.

As the smoothing process decreases accuracy, one can introduce a deviation window of ±1 residue for the agreement between each tool and MIR prediction. We also computed a consensus limited to both tools (DFIRE and I-Mutant) which is completely different from the consensus tool described in the 'Materials and methods' section. Here, it determines the average position of the minima characterized by both tools. However, if a position is found by a single tool, we keep its value to define the consensus as this position. We then compare it with MIR results also authorizing a deviation window of ±1 positions and calculate the sequence separation between the MIR and consensus positions. Table 1 summarizes all this information and the introduction of deviation in our analysis highlights six other positions of interest that do not match exactly all the conditions but are still worth

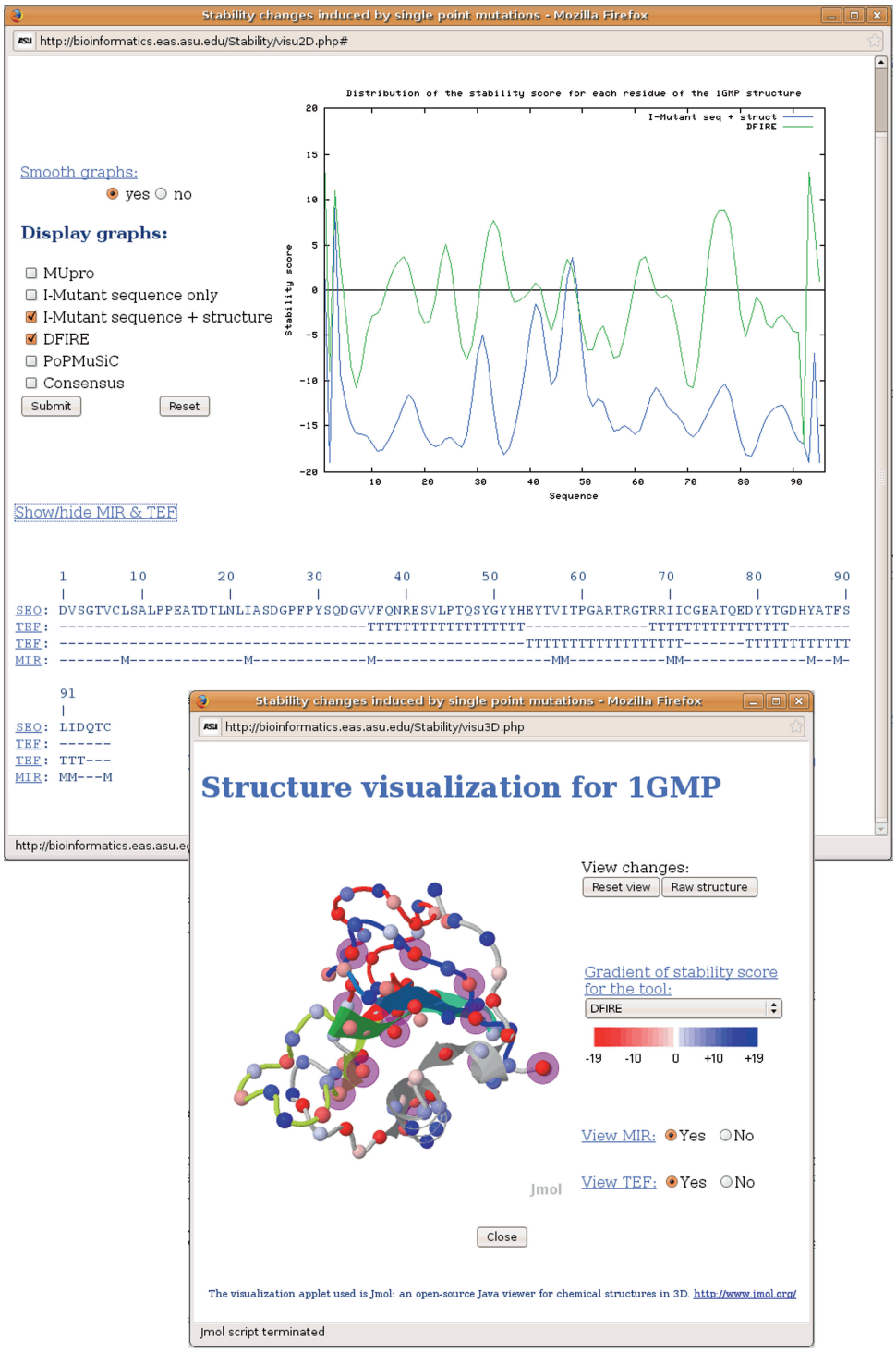


**Figure 2.** Results obtained with the 2D and 3D visualization modes for the 1gmp structure. On the upper window, the graphs correspond to smoothed stability scores along the whole sequence for the DFIRE and I–Mutant sequence + structure tools. The TEF assignment and MIR prediction are also represented below these graphs. The lower window contains a view of the 1gmp structure with the Jmol applet. The stability score for each amino acid is represented by a small sphere whose colour goes from red to blue corresponding to a score in the range of –19 to + 19. The MIR prediction is symbolized by semi–transparent purple spheres on the designated residues and the TEF assignment is characterized by the different colours on the cartoon representation.

considering. All the details regarding the validation of these positions are presented in the Supplementary Data.

The conclusion of the study case is that we can define which positions are very sensitive to mutation, participate in the maintenance of the protein structure and are highly conserved among structural families, with a degree of flexibility of one or two residues around a precise position. For 1gmp, these positions are 'around' residues: 8, 22, 36,

**Table 1.** Positions of minima of stability score for DFIRE and I-Mutant, MIR prediction, TEF assignment and solvent accessibility for the ribonuclease from *S. aureofaciens* (PDB code: 1gmp)

| | Position and related information | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DFIRE | 7 | | 20 | 28 | 37 | 44 | 52 | 56 | | | 71 | 81 | 86 | | 92 | | |
| I-Mutant | 7 | 11 | 22 | 27 | 35 | 44 | 52 | 56 | 60 | | 71 | 82 | | | 93 | | |
| Consensus | 7 | 11 | 21 | 27.5 | 36 | 44 | 52 | 56 | 60 | – | 71 | 81.5 | 86 | 92.5 | 92.5 | – | |
| MIR | 8 | | 22 | | 36 | | | 57 | 58 | 70 | 71 | | 86 | 91 | 92 | 96 | |
| Delta pos. | 1 | – | 1 | – | 0 | – | – | 1 | 2 | – | 0 | – | 0 | 1.5 | 0.5 | – | |
| TEF | 28 | | 14 | | 0 | | | 3 | 4 | 1 | 0 | | 3 | 2 | 1 | 3 | |
| RSA (%) | 12.88 | | 3.15 | | 16.90 | | | 0 | 38.34 | 0 | 0 | | 1.58 | 31.65 | 0 | 29.39 | |

The 'Consensus' line corresponds to the average stability score position for which both stability tools predict a minimum. MIR prediction is indicated, and the delta of positions correspond to the number of residues between the consensus and the nearest MIR. The sequence separation between the MIR position and the closest TEF end position is indicated in the TEF row. The RSA line (Relative Solvent Accessibility) is expressed in percentage of the amino acid surface exposed to the solvent compared to the total one.

57, 71 86, and 92. We hypothesize that mutating these amino acids would result in highly unstable protein structures and should thus be avoided.

## DISCUSSION AND PERSPECTIVES

SPROUTS development team is actively working at enhancing this first product. A recent update with about 300 new entries corresponds to roughly 64 families of at least four members, highly divergent in sequences, as the sequence identity is at most 30% between any pair of a family. From this data set a structural alignment has already been performed and the information on positions occupied only by hydrophobic residues is known (12). It has been shown that these positions statistically correspond to the folding nucleus (23). In addition to a maintenance plan to add more data in the database, a process of automatic submission for any sequence or structure is under development. We are also planning to design a decision process to guide the user in deciding which tool to use regarding the protein being studied and the type of query requested. Finally, adding other structural data such as solvent accessibility, hydrophobicity and secondary structures is considered. SPROUTS database is available at: http://bioinformatics.eas.asu.edu/sprouts.html.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

Many thanks for the authors of the different software devoted to stability changes and especially to Dr. Jean-Marc Kwasigroch for his help on using the PoPMuSiC software. We also want to acknowledge Dr. Pierre Tufféry for his help on using the RPBS resources to compute the MIR calculations.

## FUNDING

This work was partially supported by the National Science Foundation (grants IIS 0431174, IIS 0551444, and IIS 0612273) and by an invitation of the Université Pierre et Marie Curie. J.C. and M.L. benefited of an EU grant QLG2-2002-01298. Funding for open access charge: The National Science Foundation.

*Conflict of interest statement*. Any opinion, finding, and conclusion or recommendation expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

## REFERENCES

1. Fersht,A.R. (1997) Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.*, **7**, 3–9.
2. Fersht,A. and Sato,S. (2004) Φ-value analysis and the nature of protein folding transition states. *PNAS*, **101**, 7976–7981.
3. Kumar,M.D.S., Bava,K.A., Gromiha,M.M., Prabakaran,P., Kitajima,K., Uedaira,H. and Sarai,A. (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res.*, **34**, D204–D206.
4. Kawabata,T., Ota,M. and Nishikawa,K. (1999) The protein mutant database. *Nucleic Acids Res.*, **27**, 355–357.
5. Fulton,K.F., Bate,M.A., Faux,N.G., Mahmood,K., Betts,C. and Buckle,A.M. (2007) Protein Folding Database (PFD 2.0): an online environment for the International Foldeomics Consortium. *Nucleic Acids Res.*, **35**, D304–D307.
6. Papandreou,N., Berezovsky,I.N., Lopes,A., Eliopoulos,E. and Chomilier,J. (2004) Universal positions in globular proteins: observation to simulation. *Eur. J. Biochem.*, **271**, 4762–4768.
7. Miyazawa,S. and Jernigan,R.L. (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.*, **256**, 623–644.
8. Alland,C., Moreews,F., Boens,D., Carpentier,M., Chiusa,S., Lonquety,M., Renault,N., Wong,Y., Cantalloube,H., Chomilier,J. *et al.* (2005) RPBS: a web resource for structural bioinformatics. *Nucleic Acids Res.*, **33**, W44–W49.
9. Fischer,K and Marqusee,S. (2000) A rapid test for identification of autonomous folding units in proteins. *J. Mol. Biol.*, **302**, 701–712.
10. Berezovsky,I.N., Grosberg,A.Y. and Trifonov,E.N. (2000) Closed loops of nearly standard size: common basic element of protein structure. *FEBS Lett.*, **466**, 283–286.
11. Ittah,V. and Haas,E. (1995) Nonlocal interactions stabilize long range loops in the initial folding intermediates of reduced bovine pancreatic trypsin inhibitor. *Biochemistry*, **34**, 4493–4506.
12. Poupon,A. and Mornon,J.P. (1998) Populations of hydrophobic amino acids within protein globular domains; identification of conserved "topohydrophobic" positions. *Proteins*, **33**, 329–342.
13. Lamarine,M., Mornon,J.-P., Berezovsky,N. and Chomilier,J. (2001) Distribution of tightened end fragments of globular proteins

statistically match that of topohydrophobic positions: towards an efficient punctuation of protein folding? *Cell. Mol. Life Sci.*, **58**, 492–498.
14. Zhou,H. and Zhou,Y. (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.*, **11**, 2714–2726.
15. Capriotti,E., Fariselli,P. and Casadio,R. (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.*, **33**, W306–W310.
16. Cheng,J., Randall,A. and Baldi,P. (2006) Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins*, **62**, 1125–1132.
17. Gilis,D. and Rooman,M. (2000) PoPMuSiC, an algorithm for predicting protein mutant stability changes: application to prion proteins. *Protein Eng.*, **13**, 849–856.
18. Parthiban,V., Gromiha,M.M. and Schomburg,D. (2006) CUPSAT: prediction of protein stability upon point mutations. *Nucleic Acids Res.*, **34**, W239–W242.
19. Schymkowitz,J., Borg,J., Stricher,F., Nys,R., Rousseau,F. and Serrano,L. (2005) The FoldX web server: an online force field. *Nucleic Acids Res.*, **33**, W382–W388.
20. Yin,S., Ding,F. and Dokholyan,N.V. (2007) Eris: an automated estimator of protein stability. *Nat. Methods*, **4**, 466–467.
21. Masso,M. and Vaisman,I.I. (2008) Accurate prediction of stability changes in protein mutants combining machine learning with structure based computational mutagenesis. *Bioinformatics*, **24**, 2002–2009.
22. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
23. Poupon,A. and Mornon,J.P. (1999) Predicting the protein folding nucleus from sequences. *FEBS Lett.*, **452**, 283–289.