**Predict the effect of missense mutations on PTEN and TPMT protein stability**

**Qifang Xu and Roland Dunbrack, Fox Chase Cancer Center**

**Crystal Structures**

PDB structures of PTEN and TPMT are collected from PDB web site. There are five PTEN crystal structures: 1D5R, 5BUG, 5BZX, 5BZZ, 2KYL and 4O1V. 2KYL is a NMR structure with 10 models, contains coordinates from position 397 to position 403. We separated 2KYL structure into 10 structures. 4O1V is a hetero-dimer, of which chain B contains the coordinates from position 357 to position 363. All other structures contain coordinates from position 14 to position 351. There are two TPMT structures: 2BZG and 2H11. 2H11 is a homo-dimer with two TPMT chains. All TPMT structures contain coordinates from position 17 to position 245.

In order to predict the effects of missense mutations on structural stability of PTEN and TPMT proteins, we used two popular methods to predict the change of protein stability (DDG): Foldx1 and Rosetta2 . DDG of a monomeric protein induced by a point mutation is defined as the difference in free energy between the mutant structure and the wildtype structure (e.g. DDG = mutant energy – wildtype energy). The negative DDG values indicate increased stability. The effect of mutations was calculated from these DDGs, which is described as following.

**Method 1:** PositionScan of FoldX

The command PositionScan of Foldx was used to mutate each selected position to target amino acids on the monomer chains of all structures. The protein position numbers were converted into PDB residue numbers by SIFTS3. After DDG values were parsed from Foldx outputs, PDB residue numbers were converted back to protein position numbers.

Figure 1 shows the density probability plots of DDG values of each structure for both PTEN and TPMT proteins. The great majority of the mutations lied between -3 and 6 kcal/mol, which is consistent to the study of Faure et al4. Given that the free energy of protein folding is distributed roughly between 5 and 15 kcal/mol5, the effect values of mutations were calculated from prediction DDGs by the function:

The mean and standard deviation were calculated from the converted effect values over 13 monomer structures of PTEN protein and 3 monomers of TPMT protein. For those residues without coordinates, IUPred6 was used to predict their disorder probabilities. Those residues were predicted to be disordered or in a loop, N-terminus and C-terminus of the structures (Figure 2). We filled in 1.0 as effect values and 0.001 as standard deviations.

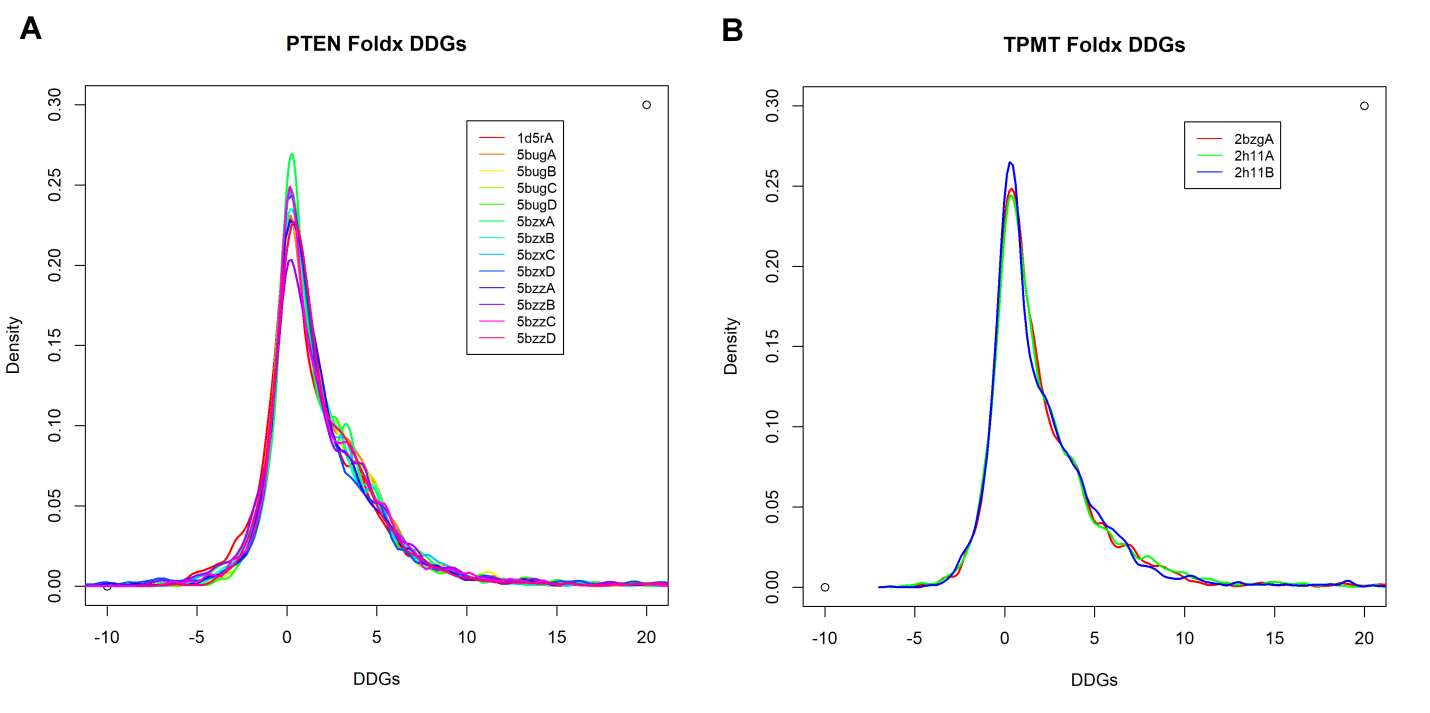


Figure 1. The density plots of PTEN and TPMT Foldx DDGs.

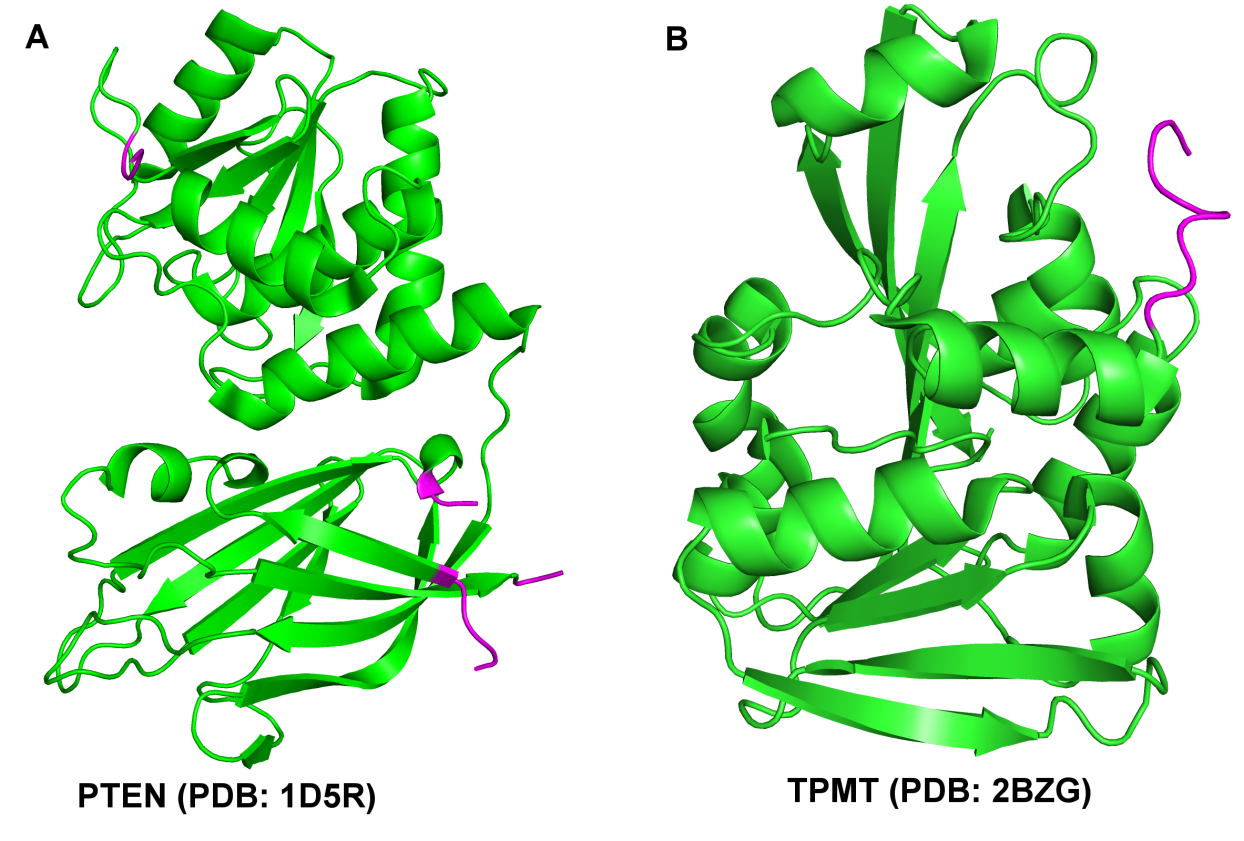


Figure 2. The structures of PTEN and TPMT proteins. The loop of missing coordinates in (A) is colored in magenta. N-terminus and C-terminus are also colored in magenta.

**Method 2 :** ddg\_monomer application in the program of Rosetta

We used the High Resolution Protocol of ddg\_monomer application (https://www.rosettacommons.org/docs/latest/application\_documentation/analysis/ddg-monomer). First, the structures were pre-minimized by minimize\_with\_cst.linuxgccrelease function, output a minimization log file for each structure. Second, a distance restraint file was generated from the minimization log file by convert\_to\_cst\_file.sh. Third, PTEN and TPMT mutation data sets were converted into Resfile format input. The parameters of ddg\_monomer were set to default settings. The number of iterations was set to 10. Fourth, we used five different methods to obtain energy scores of wildtype and mutants: (1). the average energy of 10 decoys; (2). the minimum energy; (3). the median energy. (4). the average of top-3-scoring decoys. (5). the third top scoring decoys. A DDG was calculated by subtracting wildtype energy from mutant energy. We used all five DDGs for each mutant and each structure. Last, we used the same function in “Foldx Method” to calculate the mutation effect values from DDGs. For a position, if there are 20 prediction values, we removed the bottom 5 and top 5, calculated the average and standard deviation from the middle 10 values. If there is only one structure, all five prediction values were used.

**References**

1. Schymkowitz, J.; Borg, J.; Stricher, F.; Nys, R.; Rousseau, F.; Serrano, L., The FoldX web server: an online force field. *Nucleic Acids Res* **2005,** *33* (Web Server issue), W382-8.

2. Kellogg, E. H.; Leaver-Fay, A.; Baker, D., Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins* **2011,** *79* (3), 830-8.

3. Velankar, S.; Dana, J. M.; Jacobsen, J.; van Ginkel, G.; Gane, P. J.; Luo, J.; Oldfield, T. J.; O'Donovan, C.; Martin, M. J.; Kleywegt, G. J., SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Res* **2013,** *41* (Database issue), D483-9.

4. Faure, G.; Koonin, E. V., Universal distribution of mutational effects on protein stability, uncoupling of protein robustness from sequence evolution and distinct evolutionary modes of prokaryotic and eukaryotic proteins. *Phys Biol* **2015,** *12* (3), 035001.

5. Sikosek, T.; Chan, H. S., Biophysics of protein evolution and evolutionary protein biophysics. *J R Soc Interface* **2014,** *11* (100), 20140419.

6. Dosztanyi, Z.; Csizmok, V.; Tompa, P.; Simon, I., IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **2005,** *21* (16), 3433-4.