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

Yizhou Yin [1, 2], Lipika R. Pal [1], Kunal Kundu [1, 2], John Moult [1, 3]*

[1] Institute for Bioscience and Biotechnology Research, University of Maryland, 9600 Gudelsky Drive, Rockville, MD 20850, [2] Computational Biology, Bioinformatics and Genomics, Biological Sciences Graduate Program, University of Maryland, College Park, MD 20742, USA, [3] Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD 20742.

*Corresponding author
jmoult@umd.edu
Phone: (240) 314-6241
FAX: (240) 314-6255

As a consequence of the short time line for this challenge and other activities within the group, we were able to spend only a very short time on this challenge. Thus, the approach is crude. Nevertheless, we decided to submit anyway, for the educational value.

We assumed that the experimental assay effectively measures protein abundance relative to wild type and that this quantity is closely related to the relative thermodynamic stability of proteins containing each mutation or to effects on specific biochemical mechanisms known to effect half-life such as the ubiquitination [1].

We investigated the properties of three methods of estimating the effect of mutations on thermodynamic stability with structural data: SNPs3D stability [2], Rosetta [3], and FoldX [4]. Rosetta and FoldX estimate DDG for a mutation. SNPs3D stability returns a yes/no binary estimate of whether DDG is likely greater or less than a threshold related to pathogenicity in monogenic disease, together with a confidence score. The threshold is approximately 3 Kcal/mol [2].

We also used a sequence based SVR ensemble method [5], developed in CAGI4, to estimate total activity for each mutation. The sequence based method is expected to provide an estimate of the effect on activity from all mechanisms, not just stability, and therefore to be noisy. The results were calibrated so that predictions for mutations known to have wild type or higher activities (taken from the literature and interspecies variants) had values close to 1.0, and mutations expected to have low activity (cancer drivers) have values close to 0.1.

In other work (in preparation) we have shown that a large fraction of the mutations in a protein core that decrease activity do so as a result of lowering stability. Thus, for these mutations sequence methods and stability methods should be in broad agreement. Comparison of the ensemble sequence method with both FoldX and Rosetta showed that for those mutations with a low ensemble predicted activity

there is trend to large DDG, in agreement with expectations. But overall, there is little relationship between the sequence and these two structure methods. Nevertheless, we decided to include Rosetta based predictions, largely because of good performance in a previous CAGI challenge, mutations in P16 [6]. Comparison of ensemble predicted activity and SNPs3D stability confidence scores showed an approximately linear (but noisy) relationship for predicted activities between zero and about 0.6. So we also based predictions on SNPs3D stability.

We also checked the literature for biochemical regulation of PTEN and TPMT stability [7,8,9] which assisted in deciding domain specific treatment. The PTEN ubiquitination information [1] helped to calibrate ensemble method scores.

For the ensemble method, the standard deviation for all mutations is set to the value derived in training (CAGI4). For Rosetta and SNPs3D, the standard deviations were estimated from the scatter in the comparison with the ensemble results.

The six submissions were compiled as follows:

Prediction 1: Based only on the Ensemble sequence method.

Prediction 2: Based only on the Rosetta method.

Prediction 3: Based on SNPs3D stability for mutations predicted to have an abundance between 0 and 0.6 of wild type, and on Ensemble for the rest.

Prediction 4: Based on Ensemble for core residues and on Rosetta for surface residues.

Prediction 5: For surface residues, based on SNPs3D stability for mutations predicted to have an abundance between 0 and 0.6 of wild type, and on Ensemble for the rest. For core residues, based on Ensemble.

REFERENCES

[1] Gupta, A. and Leslie, N. R. Controlling PTEN (Phosphatase and Tensin Homolog) stability: A dominant role for Lysine 66. J. Biol. Chem. 2016; 291:18465-73.

[2] Yue P, Li Z, Moult J. Loss of protein structure stability as a major causative factor in monogenic disease. J Mol Biol. 2005; 353:459-473.

[3] Das, R. and Baker, D. Macromolecular modelling with Rosetta. Annual Review of Biochemistry. 2008; 77:363-82.

[4] Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol 2002; 320:369-387.

[5] Yin, Y., Kundu, K., Pal, L. R. and Moult, J. Ensemble variant interpretation methods to predict enzyme activity and assign pathogenicity in the CAGI4 NAGLU (Human N-Acetyle Glucosaminidase) and UBE2I (Human SUMO-ligase) challenges. Human Mutation. 2017; 38:1109-22.

[6] Carraro, M., Minervini, G., …, Tosatto, S. C. E. Performance of in silico tools for the evaluation of p16INK4a (CDKN2A) variants in CAGI. Human Mutation. 2017; 38:1042-50.

[7] Leslie, N.R., Batty, I.H., …, Downes, C. P. Understanding PTEN regulation: PIP2, polarity and protein stability. Oncogene. 2008. 27: 5464-76.

[8] Rodríguez-Escudero I, Oliver, M. D.,…,Pulido R. A comprehensive functional analysis of PTEN mutations: implications in tumor- and autism-related syndromes. Hum. Mol. Genet. 2011; 20:4132-42.

[9] Wu, H., Horton, J. R., … Cheng, X. Structural basis of allele variation of human thiopurine-S-methyltransferase. Proteins. 2007; 67:198-208.