Method for the Prediction of Stability of PTEN and TPMT protein

Aditi Garg* and Debnath Pal*

E-mail: aditi18garg@gmail.com; dpal.iisc.ac.in

Method

Protein stability is a consequence of the net balance of forces. Protein stability can be estimated from energy(G)[1] and is also reflected in its flexibility(flexible regions present in a protein). While calculation of energy and its change is straight forward, it does not necessarily manifest directly in functional terms on biochemical activity and its alteration. Therefore, interpretation of flexibility graphs is an alternate paradigm one can explore for assessing single mutants for stability and functional consequences. In our case, the flexibility is calculated from the root mean square fluctuation of the structures in a Molecular Dynamics trajectory.

We ran the molecular dynamics on the C_α atoms of protein structure for 1 microsecond with **CGMM(Coarse Grained Molecular Mechanics) force field**. From the simulation, to obtain flexible regions we used the RMSF(Root Mean Square Fluctuations) values of all frames. These RMSF values are then normalized. These normalized RMSF values are real numbers so we convert them into a string as described in Bhadra *et al.*[2] On the basis of $RMSF_{norm}$ profile and the criterion for flexible region i.e. the percentage occurrence of a symbol (L $>$ 35 or combined G,H and I $>$ 14), we obtained the flexible regions of the structure. The symbols correspond to RMSF ranges of 0-1, 1-2, 2-3, $>$ 3 for L, I, H, G,

symbols, respectively.

Since the number of mutations to investigate were very high, it was not feasible for us to run all the molecular dynamics simulations. For that we ran Multiple Sequence Alignment(MSA) of the structures from Clustalx[3] with the neighbor joining method. We clustered them on the basis of phylogenetic tree obtained from the MSA so that the number of simulation is computationally feasible.Then from the cluster we took one representative and did the molecular dynamics simulation on it. As the protein structures in one cluster are very similar so we estimated the flexibility of other mutant structures from this representative structure. Only in cases where the mutation was from/to Gly or Cys, we ran all the simulations, because the coarse-grained potential function corresponding to these two residues were showing significant variations in the fluctuations. But as we did not run the simulation on other proteins in the cluster we have given higher standard deviation for those proteins in the final results.

On obtaining the molecular dynamics trajectories, the normalized RMSF and its symbolic representation (Bhadra *et al.*)[2] was used corresponding to the wild-type and the mutant. A score was derived to infer the flexibility of the mutant and wild type to estimate closeness between the two structures. For this the weighted mean value was calculated from the frequency of symbols from the flexible regions weighted by the RMSF ranges they represent, yielding a score corresponding to global flexibility estimate corresponding to the protein. A difference of the count of the symbols between the two proteins in the common flexible regions (expressed through the weighted average as defined above), gave us an estimate if the protein became more rigid/flexible compared to the wild type. The values obtained were normalized between 0-2 with a mean at 1, corresponding to the wild type.

As described above, L,I,H,G corresponds to different RMSF ranges. We counted the occurrence of each symbol in the flexible regions of wild type and mutant. Then, a difference between these numbers for each letter was calculated. On the basis of these differences the

score was calculated using the following formulae:

$$\text{Score} = \frac{l * 1 + i * 2 + h * 3 + g * 4}{l + i + h + g} \tag{1}$$

where, l = difference between occurrence of L letter in the flexible regions of wild type and mutant.

i = difference between occurrence of I letter in the flexible regions of wild type and mutant.

h = difference between occurrence of H letter in the flexible regions of wild type and mutant.

g = difference between occurrence of G letter in the flexible regions of wild type and mutant.

The global change in stability of the protein does not necessarily alter the function of the protein. There can be cases where the instability is in the distant region from the functionally important region hence not affecting the function of the protein; whereas, there can be cases where the protein is very much stable that it does not show the required amount of flexibility for proper functionality. So, in comment section we have written the distances between the ACV(auto-correlation vector) profiles of wild type and mutant flexible regions implying the difference in functionality i.e. the mutant shows the functional stability or not. This has important bearing on understanding the functional consequences of stability on the biochemical activity.

References

(1) Worth, C. L.; Preissner, R.; Blundell, T. L. *Nucleic acids research* **2011**, *39*, W215–W222.

(2) Bhadra, P.; Pal, D. *Proteins: Structure, Function, and Bioinformatics* **2014**, *82*, 2443–2454.

(3) Jeanmougin, F.; Thompson, J. D.; Gouy, M.; Higgins, D. G.; Gibson, T. J. *Trends in biochemical sciences* **1998**, *23*, 403–405.