**Analysis of mutations in Human SUMO ligase UBE2I**

**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

Consensus methods have proven very effective in CASP for prediction of protein structure [1] and prediction of the accuracy of structure models [2], but it has not so far been clear whether consensus has advantages in CAGI challenges In this challenge, we used machine learning approaches to predict UBE2I mutants’ relative cell growth rates using the output from 11 mutation analysis methods as features. A set of Human Phenylalanine hydroxylase (PAH) mutations with experimentally determined or rationally assigned enzymatic activities was used as training data. The best performing model was then used to predict relative cell growth rates for the variants in the CAGI challenge set. The experimental growth rate distribution was used to scale the prediction results in two different ways. Additional structural and functional information for UBE2I was used to adjust some predictions.

Dataset

Cell extract enzyme activity data for 92 human PAH disease related missense mutations were obtained from [3]. An additional set of 139 inter-species variants were identified by comparing the human sequence with those of seven PAH orthologs with sequence identities higher than 80%. Wild type enzyme activity was assigned to this set.

Analysis of the impact of mutations

The impact of mutations was calculated using the following eleven methods: Polyphen-2 [4], SIFT [5], SNPs3D sequence profile [6], CADD [7], Panther [8], PON-P2 [9], SNAP2 [10], Provean [11], LRT [12], VEST3 [13], MutationTaster [14]. CADD, Provean, LRT, VEST3 and MutationTaster results were extracted from the dbNSFP2.9 database [15, 16]. Results for the other methods were obtained from the corresponding web-servers. Structural stability analysis, used as described later, was obtained using standalone versions of SNPs3D stability [23] and FOLDX [24].

Training

Two sets of input features were tested: One was the set of 11 scores from the prediction methods. The other was the binary assignments (benign or deleterious) for the 11 methods. In addition, the fraction of agreement (FOA) across the methods was also included as a feature. Standard Weka [17] procedures were used for a number of machine learning methods: logistic regression, linear regression, support vector machine regression, multi-layer perceptron, M5 Rule, random tree and random forest. The best model was selected based on performance in 10-fold cross validation. The method selected was an SVM regression model with a RBF kernel with the default parameters and using the scores of the 11-methods plus the FOA as input features

Prediction

The SVR model was used to predict the expected relative enzyme activity of each challenge set mutation. For the challenge set 3, where multiple mutations are present in each sample, we assumed that the highest impact prediction dominated, and assigned that predicted activity. This assumption is supported by the observation that the experimental growth rate distribution in set 3 is strongly centered around a growth rate of zero compared to the single mutation results in sets 1 and 2. Because we trained our model on a different gene (PAH) with enzyme activity data rather than relative growth rate, we expected systematic bias in the predictions. We handled this bias by linearly scaling the prediction distributions based on the real growth rate distributions. We noted that the experimental distributions showed a number of mutations with growth rates significantly higher than wild type. In order to treat these potential gain-of-function mutations, we collected structural and functional information on UBE2I from the literature [18] and databases such as UniProt [19], PDB [20] and Cosmic [21].

Based on literature reports of up-regulation of UBE2I in cancer, we hypothesized that missense mutations increasing folding efficiency or extending *in vivo* half-life may result in higher yeast growth rates. This hypothesis is supported by two observations. First, although the overall fit of our predicted growth rate distributions to the experimental ones is reasonably good, there is an excess of predictions around wild type growth rate, particularly in set 1, suggesting that some mutations predicted as neutral are in fact cases of higher growth rate. Second, we found that the predicted neutral mutations are highly enriched with cases that FOLDX [22] predicts to increase thermodynamic stability. We therefore assigned those predicted neutral mutations with the highest predicted stability increase to higher growth rates. We also assigned higher growth rates to two mutations at positions where there are reports of increased activity in the literature [18].

Submissions

For each subset, we are submitting two versions of the predictions:

*Submission 1* has the predicted values from the support vector regression scaled to the experimental values based on comparing rankings. In addition, predicted values close to wild type that had increased stability by FOLDX were matched to the experimental values greater than one.

*Submission 2* has the predicted values scaled to the experimental distributions.

Standard deviations for the predictions are based on the root mean square error in 10-fold cross validation testing. Examination of errors as a function of activity showed no particular trend, so all errors were given the same value.

REFERENCES

1. Venclovas C, Zemla A, Fidelis K, Moult J. Assessment of progress over the CASP experiments. Proteins. 2003; 53(6):585–595.

2. Kryshtafovych A, Barbato A, Monastyrskyy B, Fidelis K, Schwede T, Tramontano A. Methods of model accuracy estimation can help selecting the best models from decoy sets: assessment of model accuracy estimations in CASP11. Proteins. 2015.

3. PAHvdb, Blau N, Yue W, Perez B, <http://www.biopku.org/pah/>

4. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nature Methods. 2010; 7:248–249.

5. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009; 4(7):1073-81.

6. Yue P, Melamud E, Moult J. SNPs3D: candidate gene and SNP selection for association studies. BMC Bioinformatics. 2006; 7:166

7. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014; 46:310-315.

8. Thomas PD and Kejariwal A. Coding single-nucleotide polymorphisms associated with complex vs Mendelian disease: Evolutionary evidence for differences in molecular effects. PNAS 2004; 101(43):15398-15403.

9. Niroula A, Urolagin S, Vihinen M. PON-P2: prediction method for fast and reliable identification of harmful variants.PLoS One. 2015; 10(2):e0117380.

10. Bromberg Y & Rost B. SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Research. 2007; 35(11): 3823-3835.

11. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics. 2015; 31(16): 2745-2747.

12. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Research. 2009; 19:1553 –1561.

13. Carter H, Douville C, Yeo G, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the Variant Effect Scoring Tool. BMC Genomics. 2013; 14(3) 1-16.

14. Schwarz JM, Cooper DN, Schuelke M, Seelow D. **MutationTaster2: mutation prediction for the deep-sequencing age.** Nat Methods. 2014; 11(4):361-2.

15. Liu X, Jian X, and Boerwinkle E. dbNSFP: a lightweight database of human non-synonymous SNPs and their functional predictions. Human Mutation. 2011; 32:894-899.

16. Liu X, Jian X, and Boerwinkle E. dbNSFP v2.0: A Database of Human Non-synonymous SNVs and Their Functional Predictions and Annotations. Human Mutation. 2013; 34:E2393-E2402.

17. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten. The WEKA Data Mining Software: An Update; SIGKDD Explorations. 2009; 11(1).

18. Bernier-Villamor V, Sampson DA, Matunis MJ, Lima CD. Structural basis for E2-mediated SUMO conjugation revealed by a complex between ubiquitin-conjugating enzyme Ubc9 and RanGAP1. Cell. 2002; 108:345–356.

19. The UniProt Consortium. UniProt: a hub for protein information. Nucleic Acids Res. 2015; 43:D204-D212.

20. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. The Protein Data Bank. Nucleic Acids Research. 2000; 28:235-242.

21. Bamford S, Dawson E, Forbes S, Clements J, Pettett R, Dogan A, et al. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. Br J Cancer. 2004; 91(2):355–8

22. 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.

23. 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.

24. 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.