SUMO ligase challenge : Mooney-Radivojac group - submission 2

Submitters: Vikas Pejaver, Chen Cui, Predrag Radivojac, Sean D. Mooney

A data set of all missense mutations was first created by taking the union of all mutations from all three subsets. MutPred2 (beta version), an algorithm for the prediction of pathogenicity of missense mutations, was then run on this data set to obtain scores between zero and one. The predictor is similar to the original MutPred[1] algorithm, except that it uses an ensemble of neural networks trained on nearly twice as many disease mutations and nearly four times as many polymorphisms. Furthermore, MutPred2 uses additional features such as the conservation of the neighborhood of the mutation and ~50 predicted structural and functional residue-level properties, among others. The underlying assumption of using this predictor is that scores of pathogenicity are expected to be correlated with the growth assay scores. A MutPred2 score of zero indicates a benign mutation (similar to wild-type) and a score of one indicates a pathogenic mutation (similar to null). In order to transform the MutPred2 score distribution to the given distribution of growth scores, the raw prediction scores were first “flipped” by subtracting each score from one. In the case of subset 3, since MutPred2 returned prediction scores for each mutation from a group of mutations individually, the mean of the flipped scores was chosen to assign a single value to each mutation group. Then, this score distribution was fitted to a beta distribution and subsequently transformed to a uniform distribution by taking the cumulative distribution function (cdf) of the beta distribution. The parameters of the distribution of experimental scores were then derived by fitting it to a Gumbel distribution. Finally, the desired distribution was obtained by evaluating the inverse cdf of this Gumbel distribution on the prediction scores from the uniform distribution. For the purposes of the challenge, all negative values were then set to zero. This procedure was carried out for each subset separately.

References:

1. Li, Biao, et al. "Automated inference of molecular mechanisms of disease from amino acid substitutions." *Bioinformatics* 25.21 (2009): 2744-2750.