**PON-P: Pathogenicity predictor of missense variants**

Abhishek Niroula and Mauno Vihinen
Department of Experimental Medical Science, Lund University, Sweden

CAGI 2012 p16 dataset submission 2

PON-P is a machine learning pathogenicity predictor of missense variants based on random forest. It predicts the probability of pathogenicity for a variation and determines the reliability of the prediction depending on which a variation is classified to one of pathogenic, neutral or unclassified variant (UV) class. The method has been trained with a variation data collected by Thusberg et al.[1] and obtained from VariBench[2].

PON-P uses random forest[3] method. A random forest is an ensemble tree based classifier that uses individual tree votes for classification. We implemented 10-fold cross validation procedure to evaluate the method performance. The training dataset was split into 10 parts using stratified random sampling. The classifier was induced on 9 parts and the remaining part was used to calculate the performance matrices. This process was repeated 10 times so that each part was employed for testing once.

To estimate the prediction reliability, the biased bootstrap estimator of standard error was implemented[4]. By stratified sampling with replacement, 200 bootstrap samples were generated from training data. Each sample was used to train a forest consisting of 300 trees. Based on the predictions of 200 classifiers, the standard error was estimated.

**CAGI p16 dataset prediction**

PON-P classifies a variation into one of pathogenic, neutral or unclassified variation classes based on random forest probability and standard deviation. The random forest probability ranges from 0 to 1 where variations are likely to be pathogenic if the probability is closer to 1 and neutral if the variations are closer to 0. For this submission, the probability and standard deviation were scaled from 0.5 to 1 using equations (1) and (2).

$$P = 0.5 \times RF_p + 0.5 \quad (1)$$

$$SD = 0.5 \times RF_{sd} \quad (2)$$

where, P is the rescaled probability of pathogenicity, $RF_p$ is the probability predicted by PON-P, SD is the rescaled standard deviation and $RF_{sd}$ is the standard deviation of PON-P.

The new version of our tool, PON-P2, is used to predict probabilities for this submission.

**References**


1. Thusberg J, Vihinen M: **Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods**. *Human mutation* 2009, **30**(5):703-714.
2. Sasidharan Nair P, Vihinen M: **VariBench: a benchmark database for variations**. *Human mutation* 2013, **34**(1):42-49.
3. Breiman L: **Random Forests**. *Machine Learning* 2001, **45**(1):5-32.
4. Sexton J, Laake P: **Standard errors for bagged and random forest estimators**. *Computational Statistics & Data Analysis* 2009, **53**(3):801-811.