Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Applications Note

OXFORD

Sequence analysis

Pathopred Web Server: deep convolutional neural network predicting pathogenicity of non-synonymous human SNPs

Corresponding Author [1,*], Co-Author [2] and Co-Author [2,*]

[1]Department, Institution, City, Post Code, Country and
[2]Department, Institution, City, Post Code, Country.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Computational tools assist in interpreting an increasing amount of data generated from large scale sequencing projects. Using novel machine learning methods and incorporating sequence information, structural information, annotations and evolutionary conservation information, high prediction accuracy of harmful amino acid substitutions in human proteins can be achieved. Trained on a VariBench benchmark dataset, a deep convolutional neural network achieves an improved prediction accuracy compared to previous methods, with an accuracy and MCC of 0.81 and 0.62 on an independent VariBench test set, respectively, when predicting the probability of pathogenic substitutions.
Availability and Implementation: The pathopred web server is freely available at http://www.pathopred.bioinfo.se/
Contact: name@bio.com
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Single nucleotide polymorphisms (SNPs) make up much of the genetic variation between humans. SNPs located in non-coding regions of the human genome can cause amino acid changes in the final protein product of genes. These non-synonymous single nucleotide polymorphisms (nsSNPs) have been found to be linked to human disorders and are documented in variation databases such as HGMD (Stenson, 2003) and dbSNP (Sherry, 2001).

With an increasing amount of variants being found and documented as sequencing technologies advance, computationally screening these variants to find those valuable for further study is important. Various computational tools exist to predict the effects of variants. The effects that these tools attempt to predict range from protein stability to the impact on transcription factor binding, to the likelihood that a variant is involved in disease.

Machine learning methods such as PON-P2 (Niroula, 2015) and PolyPhen-2 (Adzhubei, 2013) focus on the pathogenicity of nsSNPs, that is, the probability that a variant is damaging or involved in disease. These tools, and many other variant prediction tools like them, will often employ features from sequence annotations, properties of multiple sequence alignments (MSAs) constructed from protein homologues, or biochemical properties of amino acids. Certain predictors such as PON-PS (Niroula, 2017) will also attempt to predict the severity of a disease phenotype arising from an amino acid substitution.

For the purpose of assessing the performance of computational predictors, benchmark databases have been created. One such database is VariBench (Nair, 2013), containing a number of datasets with data collected and organized from dbSNP and other sources. Amongst others, it contains a dataset for protein variant tolerance, used to train PON-P2 and a disease phenotype severity annotated dataset, used to train PON-PS.

These datasets were used in (Kvist, 2018) to train a novel predictor utilising a one-dimensional deep convolutional neural network to predict both the pathogenicity as well as severity of variants in human proteins. With Pathopred, this predictor is made available through a web server.

2 Materials and methods

2.1 Datasets

The VariBench protein tolerance dataset used in (Kvist, 2018) for training the Pathopred binary pathogenicity predictor consists of a total of 28559 variants from 7675 proteins, 14674 variants of which are classed as neutral, and 13885 are classed as pathogenic. A second VariBench dataset used in (Kvist, 2018) for training the Pathopred severity predictor consists of a total of 2928 diesease variants from 91 proteins, of which 1028 variants are classes as mild, 501 as moderate, and 1399 as severe. The datasets are separated into training sets and independent test sets and are publically available at the VariBench website.

2.2 Features

The deep learning model developed in (Kvist, 2018) takes as input six different feature vectors.

For a protein of interest, A BLAST search against the NCBI nr is used to find homologous protein sequences. Two MSAs are constructed from these, one containing sequences with sequence identity below 90% to reference sequence, and one containing 90% sequence identity or above. The MSAs are constructed using MAFFT. From the mutation position, a window of 10 amino acids in each direction is used to construct three of the feature vectors.

These are the frequencies of amino acids and gap characters in the MSA along with the shannon entropy, the self-information (Hurtado, 2018) for each amino acid at a position, and the partial entropy (Hurtado, 2018) for each amino acid at a position.

Another feature vector is obtained from feeding the self-information and partial entropy into a pre-trained deep convolutional neural network (Hurtado, 2018), resulting in secondary structure, relative surface area, and representations of protein backbone torsion angles.

The amino acid sequence of the window is encoded as a one-hot vector and included as a feature vector. Finally, the original and reference amino acid are encoded as one-hot vectors, and appended to this is a log-odds score calculated from the presence of GO terms found in UniProtKB for the protein analyzed.

2.3 Predictor

The deep convolutional neural network in (Kvist, 2018) consists of several layers of one-dimensional blocks with a ResNet structure, ending in a fully-connected layer of 128 neurons with a sigmoid output layer for binary pathogenicity predictions, and a three-neuron softmax layer for severity predictions. The predictors are built using the Keras library (Chollet, 2017) and training is guided by the Adam optimizer.

3 Results and discussion

Results of the model on the independent VariBench test set for binary pathogenicity prediction are presented in Table 1, along with several others predictors evaluated in (Niroula, 2015). Pathopred improves on the other methods with an MCC of 0.62. An independent evaluation with predictors found to have high prediction accuracies in literature on the VariBench test set as well as other test sets can be seen in Figure 1. The ROC curves of Pathopred along with PON-P2 and two other predictors MVP (Qi, 2018), an ensemble method, and FATHMM (Shihab, 2013) are presented. Pathopred shows a higher AUC than both PON-P2 and FATHMM with an AUC of 0.88, but scores lower than MVP, which shows an AUC of 0.92. It should be noted that the MVP score in this comparison is likely biased because of training data overlap with test data (Kvist, 2018).

When predicting severity classes on an independent VariBench test set, Pathopred achieves an MCC of 0.21 (Kvist, 2018), as compared to PON-PS with an MCC of 0.22 (Niroula, 2017). In conclusion, while the severity prediction performs similar to earlier best performing models, the binary predictions are significantly improved.

Table 1. VariBench tolerance dataset: test set

Predictor	PPV	NPV	Sens	Spec	Acc	MCC
Pathopred	0.75	0.87	0.85	0.78	0.81	0.62
PON-P2	0.74	0.79	0.75	0.78	0.77	0.53[a]
SNAP	0.68	0.83	0.83	0.68	0.75	0.51
Condel	0.71	0.79	0.76	0.73	0.75	0.49
PolyPhen-2	0.67	0.81	0.81	0.68	0.73	0.48
SIFT	0.67	0.80	0.77	0.71	0.74	0.48
Provean	0.67	0.78	0.74	0.72	0.73	0.46

[a]Performance with all variants included, as reported in (Niroula, 2015).
Reproduced from (Kvist, 2018).

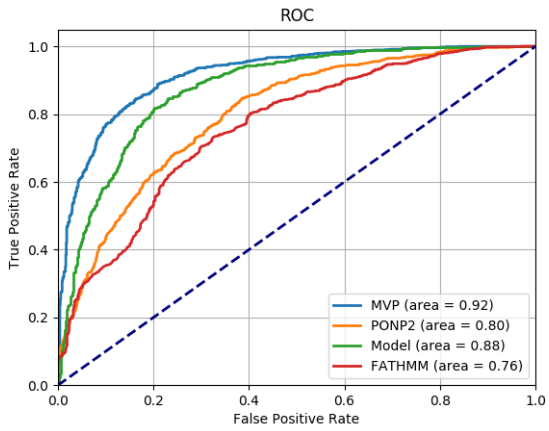

Fig. 1. ROC curves and AUC of Pathopred, MVP, FATHMM, and PON-P2.
Reproduced from (Kvist, 2018).

Funding

This work has been supported by the... Text Text Text Text.

References

Stenson, P. D., Ball, E. V., Mort, M., Phillips, A. D., Shiel, J. A., Thomas, N. S., ... and Cooper, D. N. (2003). Human gene mutation database (HGMD®): 2003 update. *Human mutation*, **21(6)**, 577-581.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic acids research*, **29(1)**, 308-311.

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

Niroula, A., and Vihinen, M. (2017). Predicting severity of disease-causing variants. *Human mutation*, **38(4)**, 357-364.

Nair, P. S., and Vihinen, M. (2013). VariBench: a benchmark database for variations. *Human mutation*, **34(1)**, 42-49.

Adzhubei, I., Jordan, D. M., and Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Current protocols in human genetics*, **76(1)**, 7-20.

Kvist, A. (2018). Identifying pathogenic amino acid substitutions in human proteins using deep learning.

Hurtado, D. M., Uziela, K., and Elofsson, A. (2018). Deep transfer learning in the assessment of the quality of protein models. arXiv preprint arXiv:1804.06281.

Chollet, F. (2017). Keras (2015).

Qi, H., Chen, C., Zhang, H., Long, J. J., Chung, W. K., Guan, Y., and Shen, Y. (2018). MVP: predicting pathogenicity of missense variants by deep learning. bioRxiv, 259390.

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L., Edwards, K. J., ... and Gaunt, T. R. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Human mutation*, **34(1)**, 57-65.