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 and dbSNP.

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 and PolyPhen-2 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 will also attempt to predict the severity of a disease phenotype arising from an amino acid substitution.

Benchmark such as VariBench systematically collect and organize datasets from databases, including dbSNP, and is used for training computational predictors and benchmarking their performance.

One such dataset contains tolerance variants, and was used for training a predictor..

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 1 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text.

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 1 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text.

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 1 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text.

Table 1. This is table caption

head1	head2	head3	head4
row1	row1	row1	row1
row2	row2	row2	row2
row3	row3	row3	row3
row4	row4	row4	row4

This is a footnote

Fig. 1. Caption, caption.

2 Materials and methods

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text.

3 Results and discussion

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text.

Table 1 shows that Text Text Text Text Text Text Text Text Text Text Text. Figure 1 shows that the above method Text Text.

Funding

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

References

Bofelli,F., Name2, Name3 (2003) Article title, *Journal Name*, **199**, 133-154.

Bag,M., Name2, Name3 (2001) Article title, *Journal Name*, **99**, 33-54.

Yoo,M.S. *et al*. (2003) Oxidative stress regulated genes in nigral dopaminergic neurnol cell: correlation with the known pathology in Parkinson's disease. *Brain Res. Mol. Brain Res.*, **110**(Suppl. 1), 76–84.

Lehmann,E.L. (1986) Chapter title. *Book Title*. Vol. 1, 2nd edn. Springer-Verlag, New York.

Crenshaw, B.,III, and Jones, W.B.,Jr (2003) The future of clinical cancer management: one tumor, one chip. *Bioinformatics*, doi:10.1093/bioinformatics/btn000.

Auhtor,A.B. *et al*. (2000) Chapter title. In Smith, A.C. (ed.), *Book Title*, 2nd edn. Publisher, Location, Vol. 1, pp. ???–???.

Bardet, G. (1920) Sur un syndrome d'obesite infantile avec polydactylie et retinite pigmentaire (contribution a l'etude des formes cliniques de l'obesite hypophysaire). PhD Thesis, name of institution, Paris, France.