doi 10.18699/bgrs2024-4.2-20

Inframe indels in human proteins: abundance and effect prediction

Ramensky V.[1, 2, 3]*

[1] *Institute for Advanced Research on Artificial Intelligence and Intelligent Systems, Lomonosov Moscow State University, Moscow, Russia*

[2] *National Medical Research Center for Therapy and Preventive Medicine of the Ministry of Healthcare of Russian Federation, Moscow, Russia*

[3] *Department of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia*

* *ramensky@gmail.com*

Key words: inframe indel; clinical significance; effect prediction; gnomAD; ClinVar

Motivation and Aim: Short insertions and deletions that do not induce frameshifts in protein (inframe indels) are of considerable interest because their functional and clinical effects remain largely unexplored and range from severe disease to no phenotypic manifestations.

Methods and Algorithms: Currently, databases contain more than 300 thousand human indels with known clinical significance and/or population frequency. Since large-scale experimental assessment of indel clinical and/or functional effects is not feasible, computational approaches that provide effect prediction are very relevant. Currently, both "classical" bioinformatics methods and novel deep learning approaches are used for effect prediction.

Results: Methods based on large protein language models (e. g., ESM1b) and not using databases of indels with known clinical relevance for training seem to be the most interesting [1], since they are free from some of the disadvantages of classical supervised learning-based methods (e. g., SIFT, CADD, Provean) [2]. We developed a method for predicting the effect of indels, based on the annotation of the sequence of the protein under study and deep alignments with homologs. We independently evaluated the performance of this approach and a number of other predictive methods [3] for indels of known clinical significance or population frequency and compared the results.

Conclusion: From the practical point of view, the most promising approach is a combination of several independent prediction methods

References

1. Brandes N., Goldman G., Wang C.H. et al. Genome-wide prediction of disease variant effects with a deep protein language model. *Nat Genet.* 2023;55:1512-1522. doi 10.1038/s41588-023-01465-0
2. Cannon S., Williams M., Gunning A.C., Wright C.F. Evaluation of in silico pathogenicity prediction tools for the classification of small in-frame indels. *BMC Med Genomics.* 2023;16:36. doi 10.1186/s12920-023-01454-6
3. Fan X., Pan H., Tian A. et al. SHINE: protein language model-based pathogenicity prediction for short inframe insertion and deletion variants. *Brief Bioinform.* 2023;24:bbac584. doi 10.1093/bib/bbac584