doi:10.1093/jmcb/mjaa007
Published online April 3, 2020

Application Note

TGPred: a tumor gene prediction webserver for analyzing structural and functional impacts of variants

With the increasing use of high-throughput sequencing technology in tumor research, a large number of somatic variations are being identified and some of them have proved to be responsible for tumorigenesis (Cancer Genome Atlas Research Network et al., 2013). Investigating structural and functional impacts of tumor somatic variants would greatly help to identify causal variations, understand the mechanisms of carcinogenesis, and develop novel anti-tumor therapies. Therefore, many efforts have recently been made to map genomic variations to 3D protein structure, such as G23D (Solomon et al., 2016) and G2S (Wang et al., 2018). Furthermore, Cancer 3D database (Porta-Pardo et al., 2015) and HotSpot3D (Niu et al., 2016) were developed to discover functional implications of mutations by means of structure data and drug information. However, there are still some limitations. Firstly, the effects of insertions and deletions (indels) are not taken into consideration. Secondly, these tools heavily depend on the resolved structures in Protein Data Bank (PDB) (Berman et al., 2000), i.e. they are not applicable when there is no reliable structural information available for wild-type protein. Here, we developed a webserver, TGPred, which provides a series of functionalities, including protein structure prediction, ligand binding site prediction, identification of functional relevant mutations, and estimation of functional impacts of mutations. Based on an interactive visualization design, these analyses are flexibly integrated, and thus the function impacts of a given protein variant could be inferred. The website is available at http://www.yylilab.cn/TGPred/.

Figure 1 shows the workflow of TGPred server. The input data consist of job ID, gene name, a DNA or protein sequence, and an amino acid (AA) variation list (see Supplementary material for format details). The input DNA sequence could be converted into a protein sequence.

Starting from a submitted or converted protein sequence, TGPred retrieves the protein structure with 100% sequence identity from PDB database in the first place; if there is no 100% sequence-identity structure available, TGPred then predicts the protein structure by using I-TASSER (Yang et al., 2015), a top-ranked approach for protein structure and function prediction. The top 10 ranked models generated by I-TASSER are adopted for the following analysis. Based on the retrieved or modeled wild-type protein structure information, the ligand binding sites could be predicted by using COACH (Yang et al., 2013), one module of I-TASSER.

TGPred allows users to define AA variations including mutations and indels. Variant protein structures involving mutations and indels could be simulated by using RASP (Miao et al., 2011) and I-TASSER, respectively.

The annotation information of cancer genomic mutations, including gene symbols, chromosome positions, transcript IDs, and AA changes and types, have been downloaded from COSMIC (Forbes et al., 2010) and reorganized as a built-in reference table that is adaptive to HotSpot3D. When users submit a gene name and AA variations, TGPred maps mutations to the reference table and extract their COSMIC annotation information. By using HotSpot3D, TGPred clusters the mutations based on their 3D spatial relationships and identifies functional relevant mutations, which have significant structural and thus functional impacts on proteins.

TGPred also estimates functional implications of mutations via PROVEAN based on sequence similarity between the submitted mutant protein sequence and the sequences from NCBI NR database (Choi and Chan, 2015). A PROVEAN score could be calculated for each mutation based on sequence evolution information. The default score threshold is −2.5, and the lower the score compared with the threshold, the more deleterious the mutation. In this way, a mutation could be classified as 'deleterious' or 'neutral'.

In TGPred, a protein structure is visualized in a user-friendly display box (Supplementary Figure S1). In order to integrate protein structural analysis to functional genomic analysis, TGPred provides an interactive visualization of analysis results. Ligand binding sites, functional relevant mutations, and indels can be highlighted in the structure model of wild-type and variant proteins. Therefore, it is feasible to observe

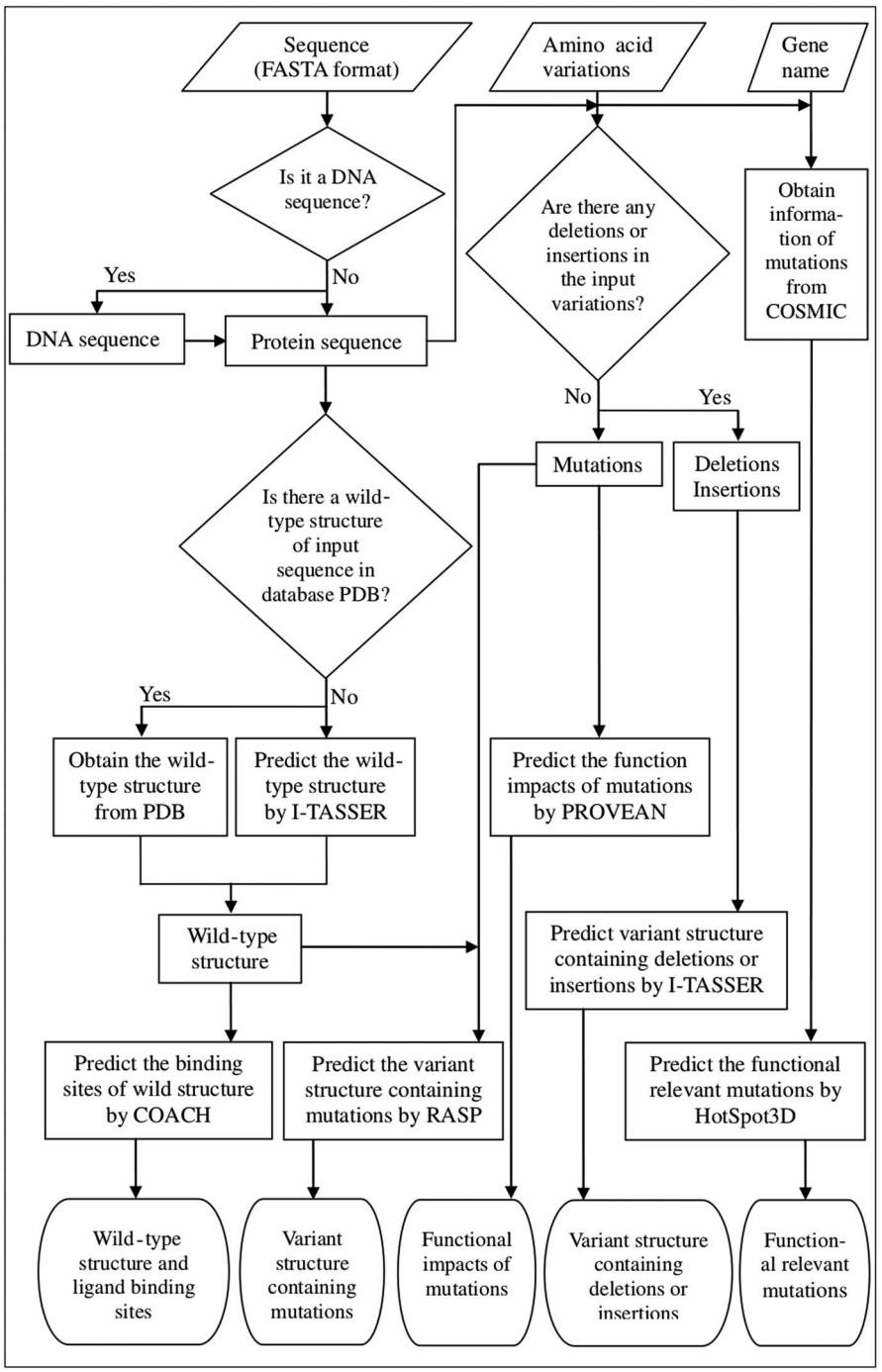

Figure 1 Workflow of TGPred server.

and investigate the protein structural alterations caused by genomic variations.

TP53 encoding p53 tumor suppressor is one of the most frequently mutated genes in human cancer (Barnoud et al., 2019) and was used as a case for TGPred. We adopted a fragment of TP53 gene as query sequence, which encodes a part of the DNA binding domain (AA101–306) of p53, consisting of 61 amino acids (AA126–186). A total of 1175 variations from COSMIC could be correlated to the query fragment, among which 10 variations (p.S127Y, p.M133K, p.F134L, p.P151S, p.G154V, p.R175H, p.C176F and p.H179R, p.P177_C182delPHHERC, and p.Y126_S127insQPHH) were taken as input variations of the query sequence. It is noted that p.R175H is the only reported 'hotspot' mutation among the 1175 COSMIC mutations and could be regarded as a spike-in control; the other nine variations were randomly selected from the 1175 COSMIC mutations.

The 3D structure of the wild-type 61-AA p53 fragment was retrieved from PDB and represented in the display box (Supplementary Figure S2). The ligand binding sites of this fragment were predicted by using COACH. The 3D structure of variant p53 fragment involving 10 input variations was simulated by RASP and I-TASSER and represented in the display box in parallel with the wild-type structure (Supplementary Figure S3A). In this way, the structural alterations caused by the variations could be easily observed and investigated. Two mutations, p.R175H and p.C176F, were identified as functionally relevant by using HotSpot3D and PROVEAN, respectively (Supplementary Figures S3B and S5). It is noticeable that both p.R175H and p.C176F are crucial mutations for the dysfunction of p53 (see Supplementary material for details of p53 analysis). We also provided BRAF example to demonstrate the webserver when protein 3D structure needs to be predicted (see Supplementary material for details).

TGPred is a user-friendly webserver developed to explore the structural and functional impacts of tumor gene variations. Compared with other analogous tools, the analysis of indels could be included in our webserver, and when wild-type protein structural information is not available, the structural and functional implications of variations still could be investigated. By integrating wild-type/variant protein structure information, ligand binding site information, spatially functional relevant mutation clustering, and functional impact estimation, TGPred provides an interactive analysis and visualization platform, which enables users to distinguish causal variations from neutral variations and understand how the variations impact cellular functions and contribute to carcinogenesis, and even helps to discover novel anti-tumor therapy targets.

[Supplementary material is available at Journal of Molecular Cell Biology online. This work was supported by grants from the National Key R&D Program of China (2018YFC0910500), the National Natural Science Foundation of China (81672736), Shanghai Municipal Science and Technology Major Project (2017SHZDZX01 and 18DZ2294200), and NIH Cancer Proteomic Tumor Analysis Consortium (CPTAC) program.]

Jixiang Liu[1], Wei Liu[1], Xue-Ling Li[1,3], Quanxue Li[1,4], Wentao Dai[1,2,*], and Yuan-Yuan Li[1,2,*]

[1]Shanghai Center for Bioinformation Technology & Shanghai Engineering Research Center of Pharmaceutical Translation, Shanghai Industrial Technology Institute, Shanghai 201203, China
[2]Department of Surgery, Shanghai Key Laboratory of Gastric Neoplasms, Shanghai Institute of Digestive Surgery, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200020, China
[3]National Engineering Research Center for Nanotechnology, Shanghai University of Medicine and Health Sciences, Shanghai 201318, China
[4]School of Biotechnology, East China University of Science and Technology, Shanghai 200237, China
*Correspondence to: Wentao Dai, E-mail: wtdai@scbit.org; Yuan-Yuan Li, E-mail: yyli@scbit.org

Edited by Luonan Chen

References

Barnoud, T., Parris, J.L.D., and Murphy, M.E. (2019). Common genetic variants in the TP53 pathway and their impact on cancer. J. Mol. Cell Biol. *11*, 578–585.

Berman, H.M., Westbrook, J., Feng, Z., et al. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.

Cancer Genome Atlas Research Network, Collisson, E.A., Mills, G.B., et al. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. *45*, 1113–1120.

Choi, Y., and Chan, A.P. (2015). PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics *31*, 2745–2747.

Forbes, S.A., Tang, G., Bindal, N., et al. (2010). COSMIC (the catalogue of somatic mutations in cancer): a resource to investigate acquired mutations in human cancer. Nucleic Acids Res. *38*, D652–D657.

Miao, Z., Cao, Y., and Jiang, T. (2011). RASP: rapid modeling of protein side chain conformations. Bioinformatics *27*, 3117–3122.

Niu, B., Scott, A.D., Sengupta, S., et al. (2016). Protein–structure-guided discovery of functional mutations across 19 cancer types. Nat. Genet. *48*, 827–837.

Porta-Pardo, E., Hrabe, T., and Godzik, A. (2015). Cancer3D: understanding cancer mutations through protein structures. Nucleic Acids Res. *43*, D968–D973.

Solomon, O., Kunik, V., Simon, A., et al. (2016). G23D: online tool for mapping and visualization of genomic variants on 3D protein structures. BMC Genomics *17*, 681.

Wang, J., Sheridan, R., Sumer, S.O., et al. (2018). G2S: a web-service for annotating genomic variants on 3D protein structures. Bioinformatics *34*, 1949–1950.

Yang, J., Roy, A., and Zhang, Y. (2013). Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. Bioinformatics *29*, 2588–2595.

Yang, J., Yan, R., Roy, A., et al. (2015). The I-TASSER suite: protein structure and function prediction. Nat. Methods *12*, 7–8.