Computational and Structural Biotechnology Journal 24 (2024) 734–738

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Analysis of mutations in precision oncology using the automated, accurate, and user-friendly web tool PredictONCO

Rayyan Tariq Khan [a,b,1], Petra Pokorna [d,f,1], Jan Stourac [a,b], Simeon Borko [a,b,c], Adam Dobias [a], Joan Planas-Iglesias [a,b], Stanislav Mazurenko [a,b], Ihor Arefiev [a], Gaspar Pinto [a,b], Veronika Szotkowska [a], Jaroslav Sterba [f,e], Jiri Damborsky [a,b], Ondrej Slaby [d,f,*], David Bednar [a,b,**]

[a] Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Brno, Czech Republic
[b] International Clinical Research Center, St. Anne's University Hospital Brno, Brno, Czech Republic
[d] Department of Biology, Faculty of Medicine and Central European Institute of Technology, Masaryk University, Brno, Czech Republic
[f] Center for Precision Medicine, University Hospital Brno, Brno, Czech Republic
[c] IT4Innovations Centre of Excellence, Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic
[e] Department of Paediatric Oncology, University Hospital Brno and Faculty of Medicine, Masaryk University, Brno, Czech Republic

ARTICLE INFO

Keywords:
Precision oncology
Webserver
Mutation
Prediction
Treatment
Next-generation sequencing
Virtual screening
Oncogenicity
Automation
Machine learning

ABSTRACT

Next-generation sequencing technology has created many new opportunities for clinical diagnostics, but it faces the challenge of functional annotation of identified mutations. Various algorithms have been developed to predict the impact of missense variants that influence oncogenic drivers. However, computational pipelines that handle biological data must integrate multiple software tools, which can add complexity and hinder non-specialist users from accessing the pipeline. Here, we have developed an online user-friendly web server tool PredictONCO that is fully automated and has a low barrier to access. The tool models the structure of the mutant protein in the first step. Next, it calculates the protein stability change, pocket level information, evolutionary conservation, and changes in ionisation of catalytic amino acid residues, and uses them as the features in the machine-learning predictor. The XGBoost-based predictor was validated on an independent subset of held-out data, demonstrating areas under the receiver operating characteristic curve (ROC) of 0.97 and 0.94, and the average precision from the precision-recall curve of 0.99 and 0.94 for structure-based and sequence-based predictions, respectively. Finally, PredictONCO calculates the docking results of small molecules approved by regulatory authorities. We demonstrate the applicability of the tool by presenting its usage for variants in two cancer-associated proteins, cellular tumour antigen p53 and fibroblast growth factor receptor FGFR1. Our free web tool will assist with the interpretation of data from next-generation sequencing and navigate treatment strategies in clinical oncology: https://loschmidt.chemi.muni.cz/predictonco/.

1. Introduction

In the last two decades, we have witnessed substantial technological advancements in human genomics, which are attributed mainly to the implementation of next-generation sequencing (NGS). With its ability to simultaneously analyse a large amount of genetic information, increasing availability, and decreasing costs, NGS has already been adopted by multiple academic and clinical laboratories and is getting to the forefront of medical diagnostics. This considerable progress and the resulting impact on clinical management is especially apparent in oncology, where comprehensive genomic profiling brings valuable information on the presence of acquired somatic alterations that can be utilised for therapeutic planning within the paradigm of precision medicine [1], with further augmentation of predictive capabilities by

* Corresponding author at: Department of Biology, Faculty of Medicine and Central European Institute of Technology, Masaryk University, Brno, Czech Republic.
** Corresponding author at: Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Brno, Czech Republic.
E-mail addresses: on.slaby@gmail.com (O. Slaby), davidbednar1208@gmail.com (D. Bednar).
[1] Joint first authors

https://doi.org/10.1016/j.csbj.2024.11.026
Received 5 August 2024; Received in revised form 12 November 2024; Accepted 12 November 2024
Available online 14 November 2024

artificial intelligence [2].

Several knowledge bases that gather published data from preclinical experiments and real-life clinical data are used to assess the potential impact of identified alterations on protein function. However, it is impossible to keep up with the amount of data generated with high-throughput technologies, and many variants lack the functional annotation necessary to distinguish oncogenic drivers from passenger events with little to no significant diagnostic, prognostic, or predictive impact. While the effect of some types of genetic variants, such as frameshift and nonsense variants, is quite definite, it is particularly challenging to predict the outcome of missense variants. This issue was soon recognized and resulted in the development of several algorithms that mainly employ information about evolutionary conservation and sequential or physicochemical properties, which might prove helpful for Mendelian disorders [3]. However, for cancer-associated proteins, a robust prediction requires a more comprehensive assessment that also employs structural data or binding properties of known inhibitors to reliably sort variants that should be pursued in preclinical studies or even clinical scenarios.

2. Minimal information for job submission

Computational pipelines that handle biological data can string together multiple software tools, each with its own settings and parameters. This can add multiple layers of complexity barring non-specialist users with little background in bioinformatics to access such a pipeline. Thus, it is important for a clinically relevant tool to have a low barrier to access. Making the tool available as an online web server would make access easier. Therefore, we have developed the new web PredictONCO, which can become a valuable tool for routine analysis of the data from next-generation sequencing experiments. The tool is designed to keep in view the urgency of oncologically relevant analysis, hence it was made fully automated, with very little input required from a user's side. Effectively, the only two pieces of information required to start a job on the web server are the target protein's name and the associated mutation. Inputting this information is done via the easy-to-use graphical user interface of the web server (Fig. 1). Once the job has started, it can take about a day to complete, but it can be longer with a load of the server. However, if the calculation for that protein and

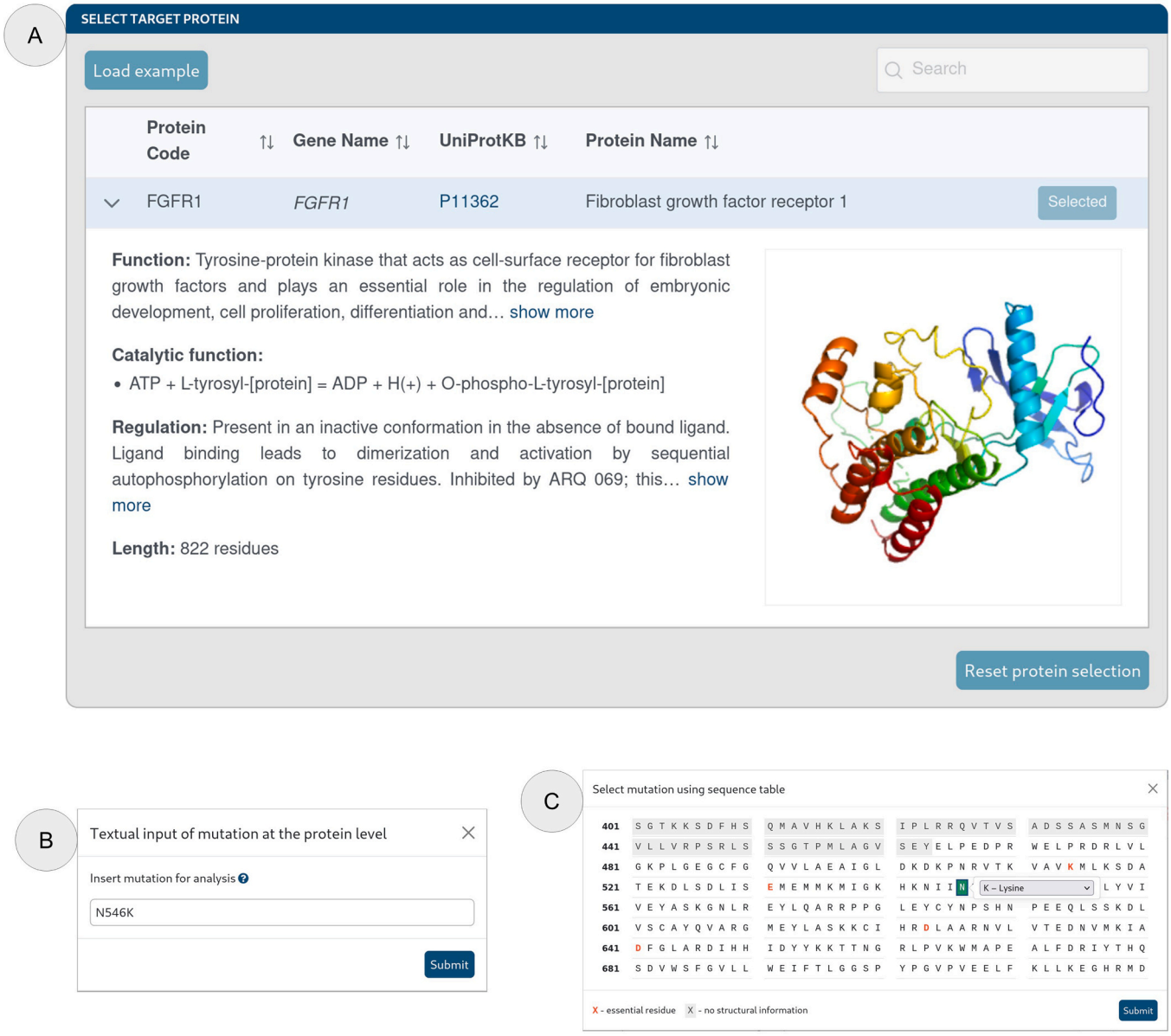

Fig. 1. The graphical user interface of PredictONCO web server's job submission page. (**A**) Protein selection window, (**B**) Mutation selection window via textual input, and (**C**) Alternative mutation selection window via the selection on the sequence table.

mutation combination has been made previously, the results are provided immediately from the jobs database. All completed calculations are added to the results page as soon as they are available, regardless of the status of the other calculations. An email alert is also sent to the user upon initiation and completion of the calculations, providing identification of the calculation and the hyperlink to the web page with results. Compared to our original study by Khan et al. [4] which contained 44 oncology-related proteins, we have updated the list for eight new targets (Supplementary Table 1). The addition of new proteins to the internal database of PredictONCO is offered to the user community based on user requests.

3. Output information and interpretation of results

The results page is an easy-to-use collection of calculations, organised in separate fields (Fig. 2). The wild type structures are used by the pipeline to calculate the stability changes using FoldX [5] and Rosetta [6]. The pipeline also models structures of the mutant proteins, and these modelled structures have pocket-level information calculated using P2Rank [7] as well as information about essential residues from

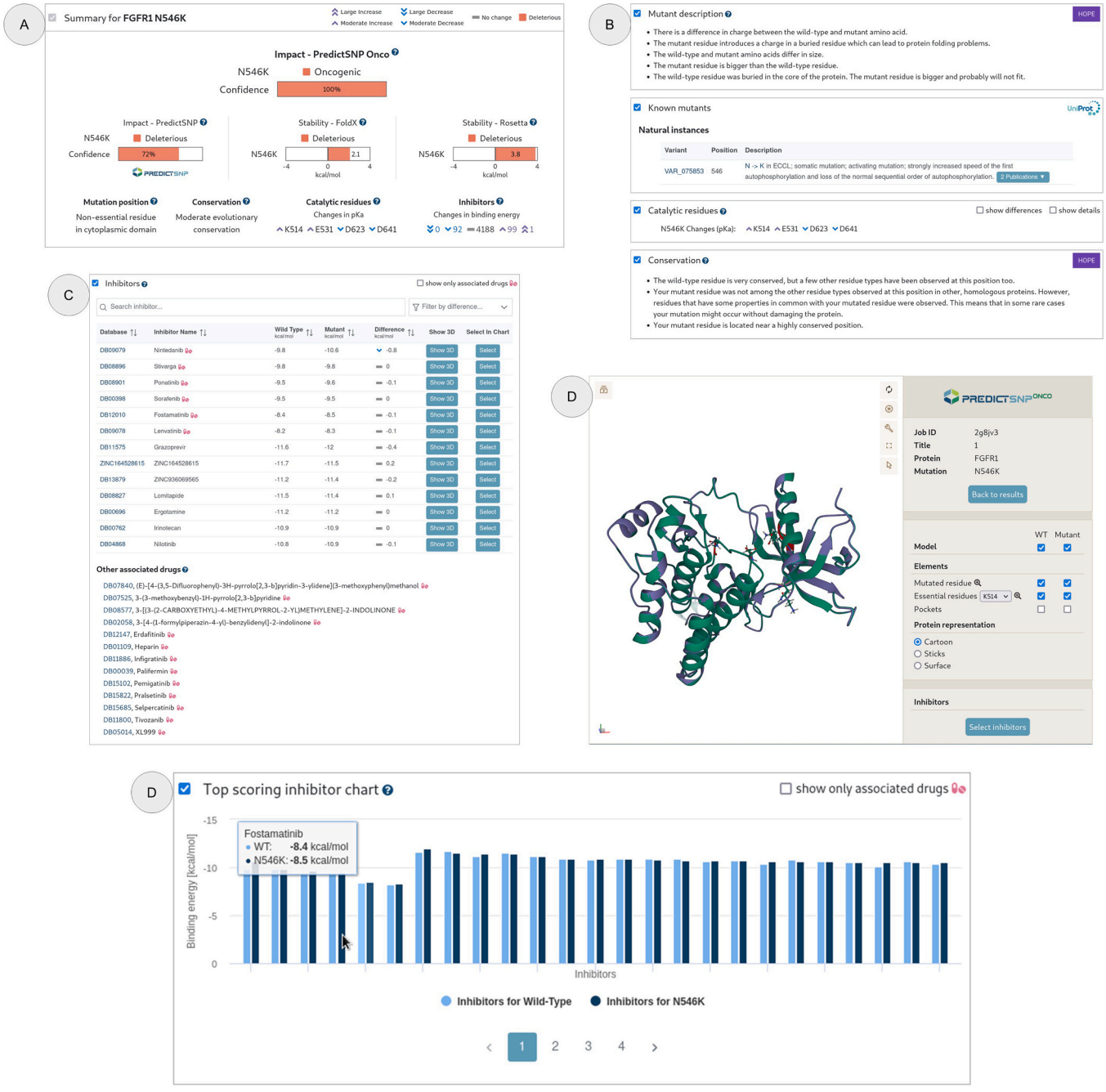

Fig. 2. The graphical user interface of PredictONCO web server's results page. (**A**) An 'at a glance' style Summary window, which compiles the most important calculations, (**B**) Various other analyses detailed in their own windows, such as Mutant description, Known mutants, and Conservation analysis from the HOPE server, as well as a window reporting changes in pKa for the catalytic residue, (**C**) Inhibitors window for showing binding energies of inhibitors in the wild type and the mutant protein, (**D**) Protein structure visualisation window for viewing the wild type and mutant protein structures with various settings. This window also allows for the visualisation of inhibitors and other protein features such as pockets and essential residues, and (**E**) The Top scoring inhibitor chart which compares the top 100 binding energies for individual inhibitors as a bar chart.

M-CSA [8] and UniProt databases [9]. The pKa values of ionizable groups, indicative of changes in reactivity between the wild type and mutant proteins are calculated using PropKa [10]. The newly developed XGBoost-based predictor uses all obtained values as features to return the probability of a mutation's oncogenic effect. To retrain the predictor with the new set of protein targets, we used an updated dataset of 592 oncogenic and 590 non-oncogenic mutations (269 new data points – Supplementary Table 2) compiled from the ClinVar [11] and OncoKB [12] databases. All mutations were annotated with a clinically verified effect based on available information from precision oncology databases [13–16] and primary literature. The predictor was validated on an independent subset of held-out data, demonstrating areas under the receiver operating characteristic curve (ROC AUC) of 0.97 and 0.94 for structure-based and sequence-based predictions, respectively (Supplementary Figure 1 and 2). The average precision from the Precision-Recall curve was 0.99 and 0.94 for structure-based and sequence-based predictions, respectively.

The results page also contains docking results of 4380 small molecules approved by the Food and Drug Administration and European Medicines Agency, docked onto both the wild type and the mutant structure using Autodock Vina [17]. Changes in the binding affinity of the drugs associated with the target protein upon mutation can support decisions on treatments (Fig. 2). The structure visualisation page allows users to inspect the tertiary structure of the wild type and mutant protein, along with the mutated residue, essential residues, and bound top-scoring drugs, in various visualisation formats. Furthermore, the results page contains information from other useful tools and databases such as UniProt [9] and HOPE [18], as well as pathogenicity scores based on the PredictSNP server [19]. The most important bits from the results page are available at the top in the 'Summary' field, along with the PredictONCO oncogenicity score. This score utilises multiple outputs of the pipeline to predict the mutation's result on the target protein's oncogenicity in a single value. To demonstrate the tool's usage, results for variants in two cancer-associated proteins, cellular tumour antigen p53, and fibroblast growth factor receptor FGFR1, are presented as case studies.

4. Case study with R175H and K139M variants of cellular tumour antigen p53

For p53, the R175H and K139M variants were submitted for analysis, with R175H being a notoriously known inactivating variant and K139M being an unknown alteration identified by comprehensive genomic profiling. Input data consisted only of the respective protein selection and selection of a particular mutation through either textual or sequence entry using a nomenclature corresponding with the canonical transcript. For R175H, the PredictONCO results showed a deleterious prediction on both the stability level and by the PredictSNP consensus classifier. Taking all calculations into account, the variant is predicted to be deleterious with a 100 % confidence score, which is in agreement with the variant being a well-known cancer-associated event leading to a loss of protein function. Its occurrence in many tumour types, of both germline and somatic origin, is also shown in the "Known mutants" section, which makes the data interpretation-related literature search easier by providing the user with relevant references.

The K139M variant of cellular tumour antigen p53, on the other hand, is a variant that lacks proper functional characterization and requires careful assessment for subsequent clinical management, especially when of germline origin, which makes it suitable for PredictONCO evaluation. PredictSNP consensus classifier predicted a deleterious effect with a moderate confidence score of 61 %, while both stability predictors predicted a neutral effect. However, differences in physicochemical properties and reported occurrence of different mutants in identical residues suggest a functional impact. The overall prediction indicates a deleterious effect with a 98 % confidence score. Therefore, by not relying just on the results of widely used sequence-based prediction algorithms, we were able to significantly increase the confidence in protein effect prediction, by 37 p.p. specifically. With such increased confidence, the clinical management of patients harbouring this germline variant would support further studies of incidence in the family and potential co-segregation with the disease.

5. Case study for N546K variant of fibroblast growth factor receptor FGFR1

A similar example can be applied to known protooncogenes with the added benefit of inhibitor binding data. Demonstrated by the example of the FGFR1 N546K variant, we got an overall deleterious prediction with a 100 % confidence score. Similar to the p53 R175H mutant, several literature references showing causality in cancers and an activating effect on protein function were available. Most importantly, as multiple inhibitors (e.g., Nintedanib, Stivarga, Ponatinib, etc.) can target FGFR1, their comprehensive overview was provided. Inhibitors were accompanied by calculated changes in binding energies, whose decreased values suggest better binding capability, which can help in the selection process if multiple options can be considered. All calculations were performed in a reasonably timely manner, under 2 h for p53 and 6 h for FGFR1, with the difference being explained by inhibitor docking and binding energy calculations.

6. Conclusions

PredictONCO is a web-based tool that uses computational algorithms to predict the effect of somatic alterations in cancer-associated proteins. It employs several algorithms and database searches that assess the impact of a mutation on protein stability, functionality, and oncogenicity. Importantly, PredictONCO also quantifies the effect of mutations on protein-drug interactions. The input for the web server is straightforward, with only the name of the target protein and associated mutation required. The results page contains several fields with different calculated properties of the mutant protein, including structure, stability change, pocket level information, and essential residues. The web server is fully automated, with email alerts sent to users upon initiation and completion of calculations, and all completed calculations are added to the results page as soon as they become available.

Web tool availability

The service PredictONCO is available free of charge to all users at the website https://loschmidt.chemi.muni.cz/predictonco/ [20].

CRediT authorship contribution statement

Petra Pokorna: Writing – review & editing, Writing – original draft, Investigation, Data curation. **Jiri Damborsky:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization. **Rayyan Tariq Khan:** Writing – review & editing, Writing – original draft, Methodology, Data curation. **Jaroslav Sterba:** Supervision, Resources, Conceptualization. **Veronika Szotkowska:** Investigation, Data curation. **Gaspar Pinto:** Writing – review & editing, Investigation, Formal analysis, Data curation. **Ihor Arefiev:** Validation, Formal analysis. **Stanislav Mazurenko:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Conceptualization. **Joan Planas-Iglesias:** Writing – review & editing, Investigation, Data curation. **Adam Dobias:** Software, Methodology. **Simeon Borko:** Writing – review & editing, Visualization, Software, Methodology. **David Bednar:** Writing – review & editing, Supervision, Software, Project administration, Funding acquisition, Conceptualization. **Jan Stourac:** Writing – review & editing, Supervision, Software, Methodology, Conceptualization. **Ondrej Slaby:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to express their thanks for the Czech Ministry of Education (CZECRIN LM2023049, RECETOX LM2023069), the Ministry of Health (NU20–03-00240) the Technology Agency of the Czech Republic (PerMed TN02000109), European Union's Horizon 2020 Research and Innovation Programme (TEAMING 857560 and 101136607), National Institute for Cancer Research (EXCELES LX22NPO5102) - Funded by the European Union - Next Generation EU, and Brno University of Technology (FIT-S-23–8209) for financial support.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.11.026.

References

[1] Azuaje F. Artificial intelligence for precision oncology: beyond patient stratification. Precis Oncol 2019;3(1):6. https://doi.org/10.1038/s41698-019-0078-1.
[2] Morash M, et al. The role of next-generation sequencing in precision medicine: a review of outcomes in oncology. J Pers Med 2018;8(3):30. https://doi.org/10.3390/jpm8030030.
[3] Alkuraya FS. Discovery of mutations for Mendelian disorders. Hum Genet 2016;135:615–23. https://doi.org/10.1007/s00439-016-1664-8.
[4] Khan RT, et al. A computational workflow for analysis of missense mutations in precision oncology. J Chemin- 2024;16:86. https://doi.org/10.1186/s13321-024-00876-3.
[5] Schymkowitz J, et al. The FoldX web server: an online force field. Nucleic Acids Res 2005;33(_2):W382–8. https://doi.org/10.1093/nar/gki387.
[6] Kellogg EH, Leaver-Fay A, Baker D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Protein: Struct, Funct, Bioinforma 2011;79(3):830–8. https://doi.org/10.1002/prot.22921.
[7] Krivák R, Hoksza D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. J Chemin- 2018;10:1–12. https://doi.org/10.1186/s13321-018-0285-8.
[8] Ribeiro AJM, et al. Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. Nucleic Acids Res 2019;46(D1):D618–23. https://doi.org/10.1093/nar/gkx1012.
[9] The UniProt Consortium, UniProt: The Universal Protein Knowledgebase in 2023. Nucleic Acids Research, 51(D1), (2023) D523–D531. https://doi.org/10.1093/nar/gkac1052.
[10] Olsson MHM, et al. PROPKA3: consistent treatment of internal and surface residues in empirical pKa predictions. J Chem Theory Comput 2011;7(2):525–37. https://doi.org/10.1021/ct100578z.
[11] Landrum MJ, et al. ClinVar: improvements to accessing data. Nucleic Acids Res 2020;48:D835–44. https://doi.org/10.1093/nar/gkz972.
[12] Chakravarty D, et al. OncoKB: a precision oncology knowledge base. JCO Precis Oncol 2017;(1):1–16. https://doi.org/10.1200/PO.17.00011.
[13] Patterson SE, et al. The clinical trial landscape in oncology and connectivity of somatic mutational profiles to targeted therapies. Hum Genom 2016;10:4. https://doi.org/10.1186/s40246-016-0061-7.
[14] Kurnit KC, et al. Personalised cancer therapy": a publicly available precision oncology resource. Cancer Res 2017;77:e123–6. https://doi.org/10.1158/0008-5472.CAN-17-0341.
[15] Gao, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 2013;6. https://doi.org/10.1126/scisignal.2004088.
[16] Ainscough BJ, et al. DoCM: a database of curated mutations in cancer. Nat Methods 2016;13:806–7. https://doi.org/10.1038/nmeth.4000.
[17] Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 2010;31(2):455–61. https://doi.org/10.1002/jcc.21334.
[18] Venselaar H, et al. Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. BMC Bioinforma 2010;11(1):1–10. https://doi.org/10.1186/1471-2105-11-548.
[19] Bendl J, et al. PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. PLoS Comput Biol 2018;10(1):e1003440. https://doi.org/10.1371/journal.pcbi.1003440.
[20] Stourac J, et al. PredictONCO: a web tool supporting decision-making in precision oncology by extending the bioinformatics predictions with advanced computing and machine learning. Brief Bioinforma 2023;25(1):bbad441. https://doi.org/10.1093/bib/bbad441.