Journal of Cheminformatics

RESEARCH

Open Access

A computational workflow for analysis of missense mutations in precision oncology

Rayyan Tariq Khan[1,3], Petra Pokorna[5,7], Jan Stourac[1,2,3], Simeon Borko[2,3,4], Ihor Arefiev[1,2], Joan Planas-Iglesias[1,2,3], Adam Dobias[1,2], Gaspar Pinto[1,2,3], Veronika Szotkowska[1,2], Jaroslav Sterba[6], Ondrej Slaby[5,7], Jiri Damborsky[1,2,3], Stanislav Mazurenko[1,2,3*] and David Bednar[1,2,3*]

Abstract

Every year, more than 19 million cancer cases are diagnosed, and this number continues to increase annually. Since standard treatment options have varying success rates for different types of cancer, understanding the biology of an individual's tumour becomes crucial, especially for cases that are difficult to treat. Personalised high-throughput profiling, using next-generation sequencing, allows for a comprehensive examination of biopsy specimens. Furthermore, the widespread use of this technology has generated a wealth of information on cancer-specific gene alterations. However, there exists a significant gap between identified alterations and their proven impact on protein function. Here, we present a bioinformatics pipeline that enables fast analysis of a missense mutation's effect on stability and function in known oncogenic proteins. This pipeline is coupled with a predictor that summarises the outputs of different tools used throughout the pipeline, providing a single probability score, achieving a balanced accuracy above 86%. The pipeline incorporates a virtual screening method to suggest potential FDA/EMA-approved drugs to be considered for treatment. We showcase three case studies to demonstrate the timely utility of this pipeline. To facilitate access and analysis of cancer-related mutations, we have packaged the pipeline as a web server, which is freely available at https://loschmidt.chemi.muni.cz/predictonco/.

Scientific contribution

This work presents a novel bioinformatics pipeline that integrates multiple computational tools to predict the effects of missense mutations on proteins of oncological interest. The pipeline uniquely combines fast protein modelling, stability prediction, and evolutionary analysis with virtual drug screening, while offering actionable insights for precision oncology. This comprehensive approach surpasses existing tools by automating the interpretation of mutations and suggesting potential treatments, thereby striving to bridge the gap between sequencing data and clinical application.

Keywords Bioinformatics, Cancer, Function, High-performance computing, Machine learning, Molecular modelling, Oncology, Personalised medicine, Single nucleotide polymorphism, Stability, Treatment

*Correspondence:
Stanislav Mazurenko
mazurenko@mail.muni.cz
David Bednar
davidbednar1208@gmail.com
Full list of author information is available at the end of the article

Introduction

More than 19 million cancer cases were diagnosed in 2020 [10] with a projected load of 28.4 million cases in 2040 [44]. The three traditionally used approaches to treat cancer, namely chemotherapy, surgery, and radiotherapy, generally result in higher mortality rates compared to the less adopted precision medicine-based techniques [27]. Next-generation sequencing technologies form the basis of precision oncology and can help generate a large amount of transcriptomic and genomic data. On the other hand, these technologies often do not provide clinically actionable data. This leads to a divide between generation of the said data and their utility, as mutants with unknown effects are often found during clinical testing [9].

There are not many tools that can help bridge the gap between data generation and creation of actionable insights. Swiss-PO, an online tool, allows for mapping experimentally determined mutations on a curated list of 50 genes and their various associated 3D structures. It also allows users to visualise multiple molecular interactions; however, it leaves it to the user to intuitively assess the structural implications of mutations that have not been experimentally determined [25] and it can also not predict patient survival outcomes. PSnpBind, a database, catalogues changes to binding affinities of ligands due to binding site single-nucleotide polymorphisms (SNPs), however this database is limited to 26 human proteins and is limited to interactions between ligands and binding site residues [2]. We sought to overcome some of these limitations by creating a robust pipeline that can predict the effects of missense mutations, even for ones which are not experimentally determined, on cancer-related proteins.

The pipeline relies on advances in fast protein modelling, such as AlphaFold [23], prediction of the effect of missense mutations on a protein structure [4], and protein stability prediction [5, 24]. This allows harvesting much more information from mutations identified by exome sequencing, which can then be used for actionable decision making. Additionally, coupling fast ligand docking in proteins [48] with the availability of multiple drug libraries online, such as ZINC [20], it is possible to screen novel potential inhibitors for the mutated proteins.

As the interpretation of large-scale genomic and transcriptomic data is limited due to the need to utilise multiple computational tools, performing the aforementioned analysis on exome sequences can take time if done manually. After a cancer diagnosis, treatment is generally a race against time, and with the variable success rates of conventional "one size fits all" therapies, fast and accurate interpretation of molecular findings and assessment of their actionability are of vital importance, especially in difficult-to-treat cases. This is where an automated precision oncology approach will be most useful as it can optimise treatment strategies, improve outcomes, and increase the quality of life for many patients [30].

Here we introduce a bioinformatics pipeline for the analysis of the effect of mutations on stability and function in cancer-related proteins. The pipeline applies in silico methods of molecular modelling, structural bioinformatics, and machine learning, and outputs actionable data which can be used for decision making. The coupled predictor produces a decision on the oncogenicity of the protein mutation by utilising the outputs derived at various stages of the pipeline. Moreover, we show the application of the pipeline on three use case studies and highlight the importance of advanced bioinformatics in precision oncology.

Methodology

Manual curation, structure repairs and geometry optimization

A list of 44 cancer-related proteins (including one isoform of a selected protein) were chosen as targets for the manual curation. The selection was based on the importance of the respective proteins for cancer diagnostics and, notably, in cancer treatment. The vast majority of curated proteins are either direct targets of therapeutic agents or, despite not being targets themselves, represent established predictive biomarkers for administering targeted treatments aimed at downstream members of the same pathway. Additionally, we included proteins that are frequently altered across various cancer types and are relevant to both diagnostics and cancer research (e.g., p53). The proteins with their various annotations are listed in the Supplementary material SI 1.

The 44 protein sequences and their annotations were fetched from the UniProt database [47]. In the case of KRAS, two isoforms are provided, including the canonical isoform and an isoform that is commonly utilized accross clinical databases of genetic variants. The essential residues were re-confirmed in the literature as well as in the Mechanism and Catalytic Site Atlas (M-CSA) [38] and the SWISS-PROT [6] databases. For the purposes of this study, in the case of multi-domain proteins, only the catalytic cytoplasmic domains of the proteins were considered. The best available structure from the wwPDB database [51], the ideal biological assembly, as well as the relevant chain (in multimeric structures) were selected based on resolution and missing parts. Canonical co-factors for structures were established using the UniProt database; these were retained in the structure, and all other ligands, ions, and water molecules were removed from the structure (SI 1). The residue indexes were mapped using the SIFTS

database [13]. After a visual inspection of each target protein, the following four key problematic regions/positions were identified: (i) missing regions, i.e., low resolution regions in the crystal structure,(ii) long, missing, and/or intrinsically disordered regions not influencing the catalytic site of the protein; (iii) missing atoms in the side chain; (iv) amino acids requiring identity correction, i.e., the sequence in the 3D structure did not correspond to that recorded in UniProt.

Each protein structure that required any of these structural improvements (for the aforementioned problematic regions/positions i, iii, or iv) was modelled using MODELLER version 9.24, 2020/04/06, r11614 [16]. The modelling was guided by the UniProt-PDB alignment provided by SIFTS. Regions identified as intrinsically disordered (repair ii) were omitted from the modelling. Custom extensions of three MODELLER Python classes (Environment, Model, and AutoModel) were developed to ensure the following: (i) the produced models incorporated any relevant co-factor from the template, (ii) the produced models were not optimised on the regions that did not require repairs, and (iii) structures containing multiple chains could be modelled and minimised at once. If no experimental structure was available, the AlphaFold database [23] was searched. The mutant structure was generated by introducing the desired mutation in the target wild type structure by MODELLER, and it was guided by a trivial alignment between the wild type and the mutant sequences.

For each protein structure, inconsistent torsion angles, total energy, or Van der Waals clashes were reduced using RepairPDB feature of FoldX 4.0 [5]. Then minimization of structures was performed in Rosetta 3.11-static [24] with constraints using the Talaris2014 force field [33]. The wild type and mutant structures were then aligned using DeepAlign 1.135-2-foss-2018b [22] to ensure that their coordinates match for further analysis.

Protein stability prediction

The impact of the missense mutation on the stability of the protein structure was calculated using Rosetta and FoldX. For FoldX the PssmStability command was used, water molecules were only taken from the 'crystal', pH was set to 7, and the number of runs was set to 5. Rosetta calculations were made on the minimised structures using the ddg_monomer command, following protocol 3 [24], for which the extent of sidechain repacking was set to within 8 Å while using the soft-rep energy function and the Talaris2014 force field.

Protein function prediction, phylogenetic analysis, and consensus classification

Additionally, PropKa 3.4.0 [40] was used to predict the impact of the mutation on the pKA values of the proteins, using the propka3 command. Homologous sequences with sufficient identity (more than 50%) and coverage (± 20% of the query sequence), i.e., UniRef50 sequences, were downloaded from the UniRef database [45], and multiple sequence alignment were generated using Clustal-Omega [42] tool from the EMBL-EBI web server [32]. This was used for conservation analysis using Jensen-Shannon Divergence algorithm [11] and transformed to mutability grades by using HotSpot Wizard [43] thresholding. The mutations were also submitted to the HOPE [49] web server to collect information from a multitude of information sources, including calculations on the 3D coordinates of the protein, sequence annotations from the UniProt database, and predictions by DAS (Distributed Annotation System services [37]. Furthermore, PredictSNP [4] was used to predict the effect of the amino acid substitution on the target protein function through consensus classification.

Pocket analysis and virtual screening

Potential binding pockets within the structures of the analysed proteins were calculated using the prank predict command in P2Rank 2.3 [26], the resulting pockets were visually analysed and manually optimised to cover the entire binding sites. Selected pockets were listed in SI 2 according to their colour codes.

Virtual screening was performed on both the wild type and the mutant protein structure. A set of 4380 small molecules that were approved by the Food and Drug Administration and European Medicines Agency was taken from the ZINC database [20]. AutoDock Vina 1.1.2 [48] was run using the standard vina command, within a parameterized grid within each protein. The grid coordinates (SI 1) were created by visually placing the grid on the protein structure in PyMOL using the ADPlugin [41] and ensuring that the binding pockets with essential residues were completely within the grid. The values for the binding energy of each small molecule to a wild type structure as well as its mutant structure were used to calculate the impact of the mutation on the binding energy.

Machine learning predictor development

The predictive part of the pipeline is a machine-learning based tool that was trained on 1073 single-point mutants whose effect was classified as Oncogenic or Benign. The variants for the Benign class were selected from the gnomAD and ClinVar [29] databases. Variants with > 1% population frequency in gnomAD, variants annotated as

"benign" or "likely benign" in the ClinVar database, and variants without ClinVar annotation, for which the classification as "benign" or "likely benign" is at the same time supported by applicable ACMG criteria [39], were utilised. The variants for the Oncogenic class were collected in expert-curated precision oncology knowledge bases, mainly, but not limited to, precision oncology knowledge base OncoKB by Memorial Sloan Kettering Cancer Center [12], as well as The JAX Clinical Knowledgebase by The Jackson Laboratory [35], Personalized Cancer Therapy Knowledge Base by MD Anderson Cancer Center [28], cBioPortal [18], and the DoCM database [1]. Variants with conflicting interpretations across multiple sources were not included in the list. Both subsets were manually filtered for any possible overlaps with the datasets used in the PredictSNP consensus predictor and its constituents.

The entire dataset (SEQ: 509 oncogenic and 564 benign data points) was further annotated by the pipeline of PredictONCO. The following six features were calculated regardless of the structural information available: essentiality of the mutated residue (yes-1/no-0), the conservation of the position (the conservation grade and MSA score), the domain where the mutation is located ("cytoplasmic", "extracellular", "transmembrane", "other"-one-hot encoded), the PredictSNP score, and the number of essential residues in the protein. For approximately half of the data (STR: 377 oncogenic and 76 benign data points), the structural information was available, and six more features were calculated: FoldX and Rosetta ddg_monomer scores, whether the residue is in the ligand-binding pocket obtained from P2Rank (yes-1/no-0), and the pKa changes of essential residues obtained from PROPKA3. The dataset is available at https://zenodo.org/records/10013764.

For the training protocol, 20% of the data in each of the two sets was kept aside for testing, chosen randomly but grouped by positions to ensure that no specific position in a protein from the test set appears in the training set. The following types of predictors were tested: the support vector machine (SVM), decision tree (DT), and XGBoost classifier (XGB), taken as they are implemented in the scikit-learn 1.2.0 and xgboost 1.7.3 libraries for Python 3.8.15. We also used the PredictSNP score alone as a baseline. For each method, we tested a set of hyperparameters based on 5-fold cross-validation implemented on the training data and receiver operating characteristic (ROC) area under the curve (AUC) as the metric (Table S1 in SI 3).

The final evaluation consisted of constructing the ROC and Precision-Recall curves. Furthermore, a round of 100 random-state re-initialisations with different random seeds was performed to evaluate the robustness of the final models. For the area under the ROC curve and the average precision values, we also reported the average and standard deviation obtained by bootstrapping (N=1000). Since any change to the predictor or data split results in a different set of x-axis coordinates in the ROC and Precision-Recall curves, we used a fixed grid of 30 points and applied 1D linear interpolation to obtain the y-axis value for each iteration. These values were then plotted as 10% and 90% quantiles.

All the training scripts, the model files, and the scripts for reproducing the model evaluations are available at https://github.com/loschmidt/predictonco-predictor/. The versions of the software tools and Python packages that were used are provided in SI 4.

Results

Development of a fully automated computational workflow

We created a bioinformatics pipeline for structure and sequence based analysis of the effects of missense mutations on cancer-related proteins (Figure S1). Since the pipeline requires curated protein structures, a method for curation was developed and applied to a list of 44 proteins (SI 1), which were then tested to ensure they can be handled in the pipeline. The pipeline was assembled using multiple bioinformatics tools, databases, and techniques. Figure 1 represents a schematic outline of the pipeline, the output of which ultimately feeds into the machine learning predictor. The predictor gives a binary decision on the effect of mutation with confidence score which is helpful in the summation and comprehension of results. Three cases of oncological interest were then studied using the developed method.

Training of sequence-based and structure-based machine learning predictors

Initially, we trained three different types of predictors, covering different trade-offs between explainability and flexibility, and compared their performance with the baseline model using the PredictSNP score alone. After optimising the hyperparameters (Table S1 in SI 3), we evaluated the performance on the held-out 20% of the dataset split by position in a protein. The support vector machines and XGBoost classifiers showed superior yet similar performance based on the area under the ROC curve and the average precision from the Precision-Recall curve (Fig. 2), also confirmed statistically (Figure S2 in SI 3). We selected the XGBoost predictor for the final model due to the interpretability of its scores: the SVM model evaluation is based on the signed distance to the separating hyperplane, without intuitive interpretation. On the other hand, the XGBoost classifier directly returns the probability that

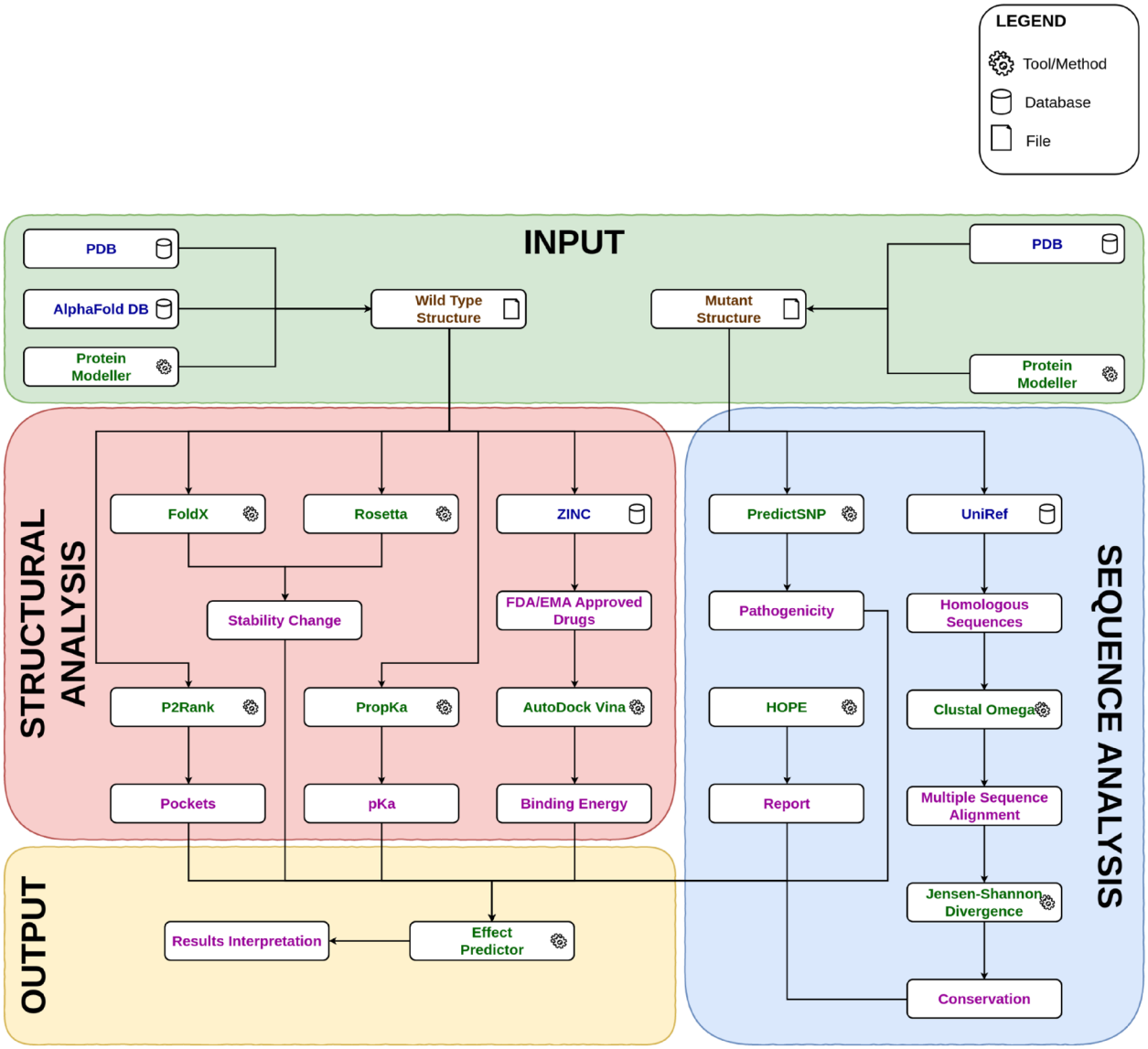

Fig. 1 A schematic representation of the bioinformatic pipeline used to predict the effect of a missense mutation on the oncogenicity of the protein

a particular mutation is oncogenic. The final XGBoost predictor is made up of 15 and 9 decision trees of the depth of 1 for structure and sequence data sets, respectively. The feature importance scores revealed that the PredictSNP score and conservation had the highest information gains (Figure S3 in SI 3). We also tested if using the train/test split by proteins would compromise the performance and saw only a marginal decrease (Figure S4 in SI 3), indicating the significant potential of the pipeline for other protein targets. The balanced accuracy for the sequence-based XGBoost predictor is 87%, and for the structure-based XGBoost predictor is 90%.

We also compared the performance of our predictor on the test set against several other models (Table 1). We evaluated the following individual scores as baselines: conservation, PredictSNP, FoldX, and Rosetta. In addition, we evaluated the performance of the ESM variants model, a recently published workflow based on the 650-million-parameter protein language ESM1b, which was used to score all possible missense variant effects in the human genome [8]. In both settings (SEQ and STR), PredictONCO showed superior performance.

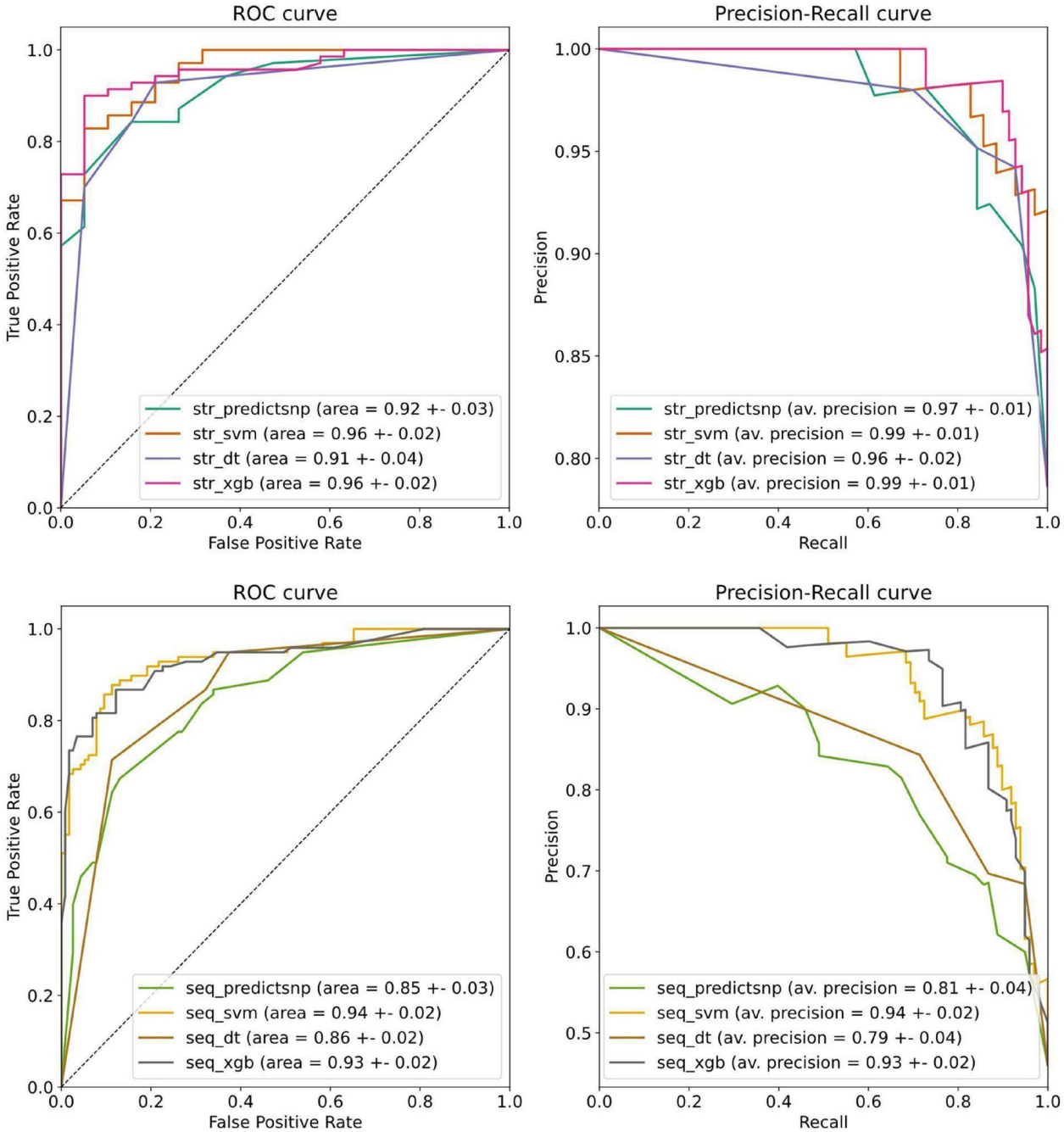

Fig. 2 The Receiver Operating Characteristic and Precision-Recall curves based on held-out test sets. Top: classifiers trained on the dataset with the structural features available (STR). Bottom: classifiers trained on the dataset with the sequence-only features (SEQ). Both the support vector machine (SVM) and XGBoost (XGB) showed comparable performance superior to the baseline model and decision tree (DT). The reported errors are standard deviations obtained by bootstrapping (N = 1000). The PredictSNP score was used as the baseline

Case studies with selected *cancer*-associated proteins

The following case studies demonstrate scenarios in which the tool has helped to facilitate further clinical decision-making. The respective variants featured in the case studies were identified across research projects utilizing high-throughput DNA sequencing techniques, which were conducted by the co-authors of this manuscript.

Table 1 Comparison of PredictONCO with other models on the test set

	Predictor	ROC AUC↑	Avg. Precision↑
SEQ	**PredictONCO**	**0.932 ± 0.018**	**0.934 ± 0.018**
	conservation	0.872 ± 0.026	0.802 ± 0.042
	predictSNP	0.845 ± 0.027	0.808 ± 0.041
	ESM variants	0.923 ± 0.018	0.911 ± 0.023
STR	**PredictONCO**	**0.955 ± 0.020**	**0.988 ± 0.006**
	FoldX	0.575 ± 0.064	0.867 ± 0.037
	Rosetta	0.628 ± 0.064	0.876 ± 0.039
	conservation	0.937 ± 0.037	0.970 ± 0.020
	predictSNP	0.918 ± 0.030	0.973 ± 0.011
	ESM variants	0.929 ± 0.027	0.981 ± 0.009

PredictONCO values are in bold

The models selected for comparison were individual features and the ESM variants predictor. The reported errors are standard deviations obtained by bootstrapping (N = 1000).

Case study 1-platelet derived growth factor receptor *beta* PDGFRB N666T

In a patient with myofibroma, sequencing analysis revealed an N666T variant of the PDGFRB protein (UniProt ID: P09619). Even though some mutations of the N666 residue, including N666K [21], N666H [36], or N666S [34], have already been documented in myofibroma patients, N666T, in particular, lacks published functional evidence and was reported in a total of one patient in combination with another mutation. Therefore, a comprehensive assessment of its effect would provide further confirmatory evidence on the variant's pathogenicity, which is substantial, given the therapeutic implications of receptor tyrosine kinase inhibition. Conservation status showed high evolutionary conservation of mutated position. For amino acid 826, one of the essential catalytic residues, a large increase in dissociation constant was predicted, suggesting a significant functional impact. Both stability predictors suggested a deleterious effect, which is also in agreement with the deleterious effect on protein function predicted by PredictSNP. Given all this data, the oncogenic effect was predicted by the XGBoost classifier with 100% confidence. Furthermore, in virtual screening, Sunitinib showed a slightly better increase in binding affinity compared to Imatinib, which was used as a drug of choice in different myofibroma preclinical studies, making Sunitinib a suitable alternative option for therapeutic planning. The full report can be accessed at https://loschmidt.chemi.muni.cz/predictonco/job/pdgfrb_N666T

Case study 2-angiopoietin-1 receptor TIE2 G1036D

In a patient with a vascular tumour, sequencing analysis revealed a G1036D variant in the TIE2 (UniProt ID: Q02763) gene. The G1036D variant represents a previously undescribed alteration, which has not been documented in the literature, clinical, or population databases of genetic variants. Given the rapidly evolving field of vascular tumour genetics and the possibility of targeted therapeutics administration, identifying novel potentially activating alterations is vastly important. Although the residue is non-essential, moderately evolutionarily conserved, and only moderate changes were predicted for the catalytic residues, the overall impact was evaluated by the XGBoost classifier as oncogenic with a 99% confidence score and was based on a deleterious prediction by both the PredictSNP algorithm and stability predictors FoldX and Rosetta. This could be approached as a basis to facilitate further functional tests to measure mutant receptor phosphorylation and, if proven as activating, introduce a considerable therapeutic opportunity (by potentially using one of the suggested inhibitive compounds such as Ecteinascidin, Ponatinib, etc., or other inhibitors of downstream signalling cascade) as well as an addition to the knowledge on disease pathogenesis. The full report can be accessed at https://loschmidt.chemi.muni.cz/predictonco/job/tie2_G1036D

Case study 3-tumour protein p53 K101Q

In next-generation sequencing screening for cancer predispositions, the K101Q variant of p53 (UniProt ID: P04637) was identified in an individual with a negative family history of cancer. p53 represents the most commonly altered gene in all cancers, and p53 variants predispose to cancer development when of germline origin. Therefore, a careful assessment must be performed for further genetic counselling. The respective variant has not been documented in the literature or functionally characterised. With lacking evidence from literature and databases of genetic variants, typically only prediction algorithms that employ sequence-based information without structural data are available. Therefore, combining both structural and sequence-related perspectives might yield a more accurate prediction. The XGBoost classifier predicted the mutation as neutral with an 81% confidence score, supported by both the PredictSNP prediction and the stability predictors. Information on evolutionary conservation showed that the wild-type residue is not conserved at this position, which may suggest that the variant is not damaging to the protein. Based on these results and no family history of cancer, the variant should not influence subsequent clinical management. Given

the importance of p53 variants in both somatic and germline contexts and their same functional impact, this case study exemplifies the utility of the tool in the assessment of hereditary cancer predisposition. The full report can be accessed at the following link—https://loschmidt.chemi.muni.cz/predictonco/job/p53_K101Q

Discussion

Prediction of the effect of missense mutations on cancer-related protein structures is a complicated task. This paper presents our pipeline for tackling this problem, thus allowing clinical bioinformaticians to easily run multiple cancer-related analyses for their target mutations on a curated list of proteins.

A major part of the pipeline capitalises on structural bioinformatics, and it requires the presence of good quality protein structures for accurate analysis. However, a high number of cancer-associated structures are transmembrane channels and thus only have fragmented domain-level structures. Some of them can be multimeric, and thus modelling proves a challenge. Despite AlphaFold [23] being touted as a major groundbreaker in the field of protein structure modelling, it proves inefficient in modelling large multi-subunit, multimeric proteins as quaternary domain level interactions are difficult to model. Thus the structural bioinformatics part of the pipeline is limited to working with high-quality structures at the domain level. AlphaFold-Multimer [17] can be used to predict the multimeric conformation in 70% of heteromeric cases and 72% of homomeric cases to limit this problem, and it is unclear whether this accuracy of predictions is viable for working with oncogenic or tumour suppressor proteins, especially when the final prediction will likely be used in a medical context.

Currently, the web server provides predictions for 44 target proteins, which were selected based on their relevance to the field of oncology. Appropriate processing of a new structure to be used in the pipeline requires expert-level knowledge of multiple bioinformatic tools. Curation in this field is a recognized bottleneck, especially in the case of the interpretation of results [7].) The addition of new target proteins to the internal database connected to the PredictONCO web server is possible and it is offered to the user community based on direct requests. Once a protein is curated, all mutations in its structure can be easily analysed. Moreover, the pipeline can also work with sequence-only data, and the trained XGBoost classifier can also reliably predict using only the sequence-based features, with only a 4% drop in average precision.

The pipeline has no standard run time as it mostly depends on whether structural analysis needs to be performed along with sequence-based analysis or not. The structural analysis increases the computational load, and the complexity of the structure can further increase the run time. However, the calculations generally do not take more than two days to complete. It is unclear whether this time frame is long or short as run time benchmarking would require the existence of other similar tools, techniques or pipelines for comparative purposes, and specialised methodologies that deal with the same case do not exist. However, this time window meets the initial requirements for the use of the web server in clinical practice as well as for research and educational purposes. Furthermore, it helps assist in making the result interpretation step easier as interpretation itself is a recognized bottleneck [7].)

Comparison to other similar tools is difficult as, as of this writing, we did not come across a pipeline integrating multiple approaches to predict the effect of a missense mutation on a cancer-related protein. However several databases and online data integrating tools do exist. The two most prominent of these databases are the International Cancer Genome Consortium (ICGC) [46] and The Cancer Genome Atlas (TCGA) [50]. Furthermore, survival analysis tools also exist and are primarily based on 4 types of data: (i) mRNA data, such as PRECOG [19], (ii) ncRNA data, such as OncoLnc [3], (iii) DNA methylation and mutation data, such as cBioPortal [18], and (iv) Protein data, such as TCPA [31]. Additionally, the Swiss-PO web tool for mapping gene mutations on the 3D structure can be used, but it only allows for intuitive and qualitative analysis of mutations that have already been experimentally determined [25]. In comparison to the aforementioned database, PSnpBind is also difficult as it only catalogues changes to binding affinities of ligands due to binding site single-nucleotide polymorphisms (SNPs) [2].

Our pipeline currently only supports missense mutations, as it is unable to handle insertions, deletions, or fusions of oncogenic proteins because individual tools in the pipeline are not able to analyse them. However, substitutions do make up a large number of cancer-associated mutations as a large number of genes associated with various cancer types contain single nucleotide variants [15]. For common solid tumours, 95% of cancer driver mutations in humans are single-base substitutions. Approximately, 90.7% of these result in the amino acid being substituted for another, non-synonymous one [14]. Thus, even though insertions, deletions, and fusions cannot be analysed using the pipeline, it still provides predictions for a significant majority of cancer-related alterations. The tool is freely accessible to the community of bioinformaticians and

medical doctors and will provide fast and useful analysis of data from the sequencing of patient samples.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-024-00876-3.

Supplementary Material 1.

Supplementary Material 2.

Supplementary Material 3.

Author contributions

Conceptualization: JSte, OS, JD, SM, DB Methodology: RTK, JS, JPI, GP, JD, SM, DB Data analysis: RTK, PP, JS, SB, IA, JPI, VS, SM, DB Software development: JS, SB, JPI, AD Writing the main manuscript: RTK, PP, JS Supervision: JSte, OS, JD, SM, DB All authors contributed to the final version of the manuscript.

Funding

The authors would like to express their thanks to the Czech Ministry of Education [ESFRI CZECRIN LM2023049; ESFRI eINFRA LM2018140, ESFRI RECETOX LM2023069]; the Technology Agency of the Czech Republic [TREND FW03010208; PERMED TN02000109]; the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 857560 (CETOCOEN Excellence); Brno University of Technology [FIT-S-23-8209]; Ministry of Health [NU20-03-00240]. The research was further supported by the project National Institute for Oncology Research [Programme EXCELES, ID Project No. LX22NPO5102 funded by the European Union—Next Generation EU]. This publication reflects only the author's view, and the European Commission is not responsible for any use that may be made of the information it contains.

Availability of data and materials

The pipeline is available as a web server, at https://loschmidt.chemi.muni.cz/predictonco/. The list of proteins, definition of binding pockets, and ML model validation are attached as supplementary files. The training and testing datasets are available at https://zenodo.org/records/10013764.

Declarations

Competing interests

None declared.

Author details

[1]Loschmidt Laboratories, Department of Experimental Biology, Faculty of Science, Masaryk University, Brno, Czech Republic. [2]Loschmidt Laboratories, RECETOX, Faculty of Science, Masaryk University, Brno, Czech Republic. [3]International Clinical Research Center, St. Anne's University Hospital Brno, Brno, Czech Republic. [4]IT4Innovations Centre of Excellence, Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic. [5]Central European Institute of Technology, Masaryk University, Brno, Czech Republic. [6]Department of Paediatric Oncology, University Hospital Brno and Faculty of Medicine, Masaryk University, Brno, Czech Republic. [7]Department of Biology, Faculty of Medicine, Masaryk University, Brno, Czech Republic.

Received: 22 November 2023 Accepted: 26 June 2024
Published online: 29 July 2024

References

1. Ainscough BJ et al (2016) DoCM: a database of curated mutations in cancer. Nat Method 13(10):806–807. https://doi.org/10.1038/nmeth.4000
2. Ammar A et al (2022) PSnpBind: a database of mutated binding site protein–ligand complexes constructed using a multithreaded virtual screening workflow. J Chemin. https://doi.org/10.1186/s13321-021-00573-5
3. Anaya J (2016) OncoLnc: linking TCGA survival data to MRNAs, MiRNAs, and LncRNAs. PeerJ Comput Sci 2:e67. https://doi.org/10.7717/peerj-cs.67
4. Bendl J et al (2014) PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. PLoS Comput Biol. https://doi.org/10.1371/journal.pcbi.1003440
5. Blanco JD et al (2018) FoldX accurate structural protein–DNA binding prediction using PADA1 (protein assisted DNA assembly 1). Nucl Acid Res 46(8):3852–3863. https://doi.org/10.1093/nar/gky228
6. Boeckmann B (2003) The SWISS-PROT protein knowledgebase and its Supplement TrEMBL in 2003. Nucl Acid Res 31(1):365–370. https://doi.org/10.1093/nar/gkg095
7. Bungartz KD et al (2018) Making the right calls in precision oncology. Nat Biotechnol 36(8):692–696. https://doi.org/10.1038/nbt.4214
8. Brandes N et al (2023) Genome-wide prediction of disease variant effects with a deep protein language model. Nat Genet. https://doi.org/10.1038/s41588-023-01465-0
9. Buzdin A et al (2021) Editorial: next generation sequencing based diagnostic approaches in clinical oncology. Front Oncol. https://doi.org/10.3389/fonc.2020.635555
10. "Cancer Today." *Iarc.fr*, 2020, https://gco.iarc.fr/today/home.
11. Capra JA, Singh M (2007) Predicting functionally important residues from sequence conservation. Bioinformatics 23(15):1875–1882. https://doi.org/10.1093/bioinformatics/btm270
12. Chakravarty D et al (2017) OncoKB: a precision oncology knowledge base. JCO Precis Oncol. https://doi.org/10.1200/po.17.00011
13. Dana JM et al (2018) SIFTS: updated structure integration with function, taxonomy and sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. Nucl Acid Res. https://doi.org/10.1093/nar/gky1114
14. Darbyshire M et al (2019) Estimating the frequency of single point driver mutations across common solid tumours. Sci Rep. https://doi.org/10.1038/s41598-019-48765-2
15. Deng N et al (2017) Single nucleotide polymorphisms and cancer susceptibility. Oncotarget. https://doi.org/10.1632/oncotarget.22372
16. Eswar N et al (2008) Protein structure modeling with MODELLER. Method Mol Biol. https://doi.org/10.1007/978-1-60327-058-8_8
17. Evans R et al (2021) Protein complex prediction with AlphaFold-Multimer. BioRxiv. https://doi.org/10.1101/2021.10.04.463034
18. Gao J et al (2013) Integrative analysis of complex cancer genomics and clinical profiles using the CBioPortal. Sci Signal. https://doi.org/10.1126/scisignal.2004088
19. Gentles AJ et al (2015) The prognostic landscape of genes and infiltrating immune cells across human cancers. Nat Med. https://doi.org/10.1038/nm.3909
20. Irwin JJ et al (2012) ZINC: a free tool to discover chemistry for biology. J Chem Inf Model. https://doi.org/10.1021/ci3001277
21. Iwamura R et al (2023) PDGFRB and NOTCH3 mutations are detectable in a wider range of pericytic tumors, including myopericytomas, angioleiomyomas, glomus tumors, and their combined tumors. Mod Pathol. https://doi.org/10.1016/j.modpat.2022.100070
22. Jiménez-Moreno A et al (2021) DeepAlign, a 3D alignment method based on regionalized deep learning for Cryo-EM. J Struct Biol 213(2):107712. https://doi.org/10.1016/j.jsb.2021.107712
23. Jumper J et al (2021) Highly accurate protein structure prediction with AlphaFold. Nature. https://doi.org/10.1038/s41586-021-03819-2
24. Kellogg EH et al (2010) Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Protein Struct Funct Bioinform 79(3):830–838. https://doi.org/10.1002/prot.22921
25. Krebs FS et al (2021) Swiss-PO: a new tool to analyze the impact of mutations on protein three-dimensional structures for precision oncology. NPJ Precis Oncol 5(1):19. https://doi.org/10.1038/s41698-021-00156-5
26. Krivák R, Hoksza D (2018) P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. J Cheminformatics. https://doi.org/10.1186/s13321-018-0285-8
27. Krzyszczyk P et al (2018) The growing role of precision and personalized medicine for cancer treatment. Technology. https://doi.org/10.1142/s2339547818300020
28. Kurnit KC et al (2017) 'Personalized cancer therapy': a publicly available precision oncology resource. Cancer Res 77(21):e123–e126. https://doi.org/10.1158/0008-5472.can-17-0341

29. Landrum MJ et al (2017) ClinVar: improving access to variant interpretations and supporting evidence. Nucl Acid Res 46(D1):D1062–D1067. https://doi.org/10.1093/nar/gkx1153
30. Lassen UN et al (2021) Precision oncology: a clinical and patient perspective. Futur Oncol 17(30):3995–4009. https://doi.org/10.2217/fon-2021-0688
31. Li J et al (2013) TCPA: a resource for cancer functional proteomics data. Nat Method 10(11):1046–1047. https://doi.org/10.1038/nmeth.2650
32. Madeira F et al (2022) Search and sequence analysis tools services from EMBL-EBI in 2022. Nucl Acid Res 50(W1):W276–W279. https://doi.org/10.1093/nar/gkac240
33. O'Meara MJ et al (2015) Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with rosetta. J Chem Theor Computation. 11(2):609–622. https://doi.org/10.1021/ct500864r
34. Ortiz E et al (2020) Invasive myofibromatosis with visceral involvement in a term newborn: a case report. Am J Pediatr 6(2):173–173. https://doi.org/10.11648/j.ajp.20200602.30
35. Patterson SE et al (2016) The clinical trial landscape in oncology and connectivity of somatic mutational profiles to targeted therapies. Hum Genom. https://doi.org/10.1186/s40246-016-0061-7
36. Pond D et al (2018) A patient with germ-line gain-of-function PDGFRB P.N666H mutation and marked clinical response to imatinib. Genet Med 20(1):142–150. https://doi.org/10.1038/gim.2017.104
37. Prlić A et al (2007) Integrating sequence and structural biology with DAS. BMC Bioinform. https://doi.org/10.1186/1471-2105-8-333
38. Ribeiro AJM et al (2017) Mechanism and catalytic site atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. Nucl Acid Res 46(D1):D618–D623. https://doi.org/10.1093/nar/gkx1012
39. Richards S et al (2015) Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. Genet Med 17(5):405–424. https://doi.org/10.1038/gim.2015.30
40. Rostkowski M et al (2011) Graphical analysis of PH-dependent properties of proteins predicted using PROPKA. BMC Struct Biol. https://doi.org/10.1186/1472-6807-11-6
41. Seeliger D, de Groot BL (2010) Ligand docking and binding site analysis with PyMOL and autodock/vina. J Comput Aided Mol Des 24(5):417–422. https://doi.org/10.1007/s10822-010-9352-6
42. Sievers F et al (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. Mol Syst Biol 7(1):539. https://doi.org/10.1038/msb.2011.75
43. Sumbalova L et al (2018) HotSpot wizard 3.0: web server for automated design of mutations and smart libraries based on sequence input information. Nucl Acid Res 46(W1):W356–W362. https://doi.org/10.1093/nar/gky417
44. Sung H et al (2021) Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA A Cancer J Clin 71(3):209–249. https://doi.org/10.3322/caac.21660
45. Suzek BE et al (2014) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 31(6):926–932. https://doi.org/10.1093/bioinformatics/btu739
46. The International Cancer Genome Consortium (2010) International network of cancer genome projects. Nature 464(7291):993–998. https://doi.org/10.1038/nature08987
47. The UniProt Consortium (2022) UniProt: the universal protein knowledgebase in 2023. Nucl Acid Res 51(D1):D523-531. https://doi.org/10.1093/nar/gkac1052
48. Trott O, Olson AJ (2009) AutoDock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. https://doi.org/10.1002/jcc.21334
49. Venselaar H et al (2010) Protein structure analysis of mutations causing inheritable diseases. An e-science approach with life scientist friendly interfaces. BMC Bioinform. https://doi.org/10.1186/1471-2105-11-548
50. Weinstein JN et al (2013) The cancer genome atlas pan-cancer analysis project. Nat Genet 45(10):1113–1120. https://doi.org/10.1038/ng.2764
51. wwPDB Consortium (2018) Protein data bank: the single global archive for 3D macromolecular structure data. Nucl Acid Res 47(D1):D520–D528. https://doi.org/10.1093/nar/gky949

Publisher's Note