Babbi et al. Top Italian Scientists Journal, 2024;1(2)
https://doi.org/10.62684/HDYO1899

Top Italian Scientists Journal

Open Access | Article | Biomedical Sciences

ISSN 3033-5132

Bioinformatic Sweeties: a unified portal for characterizing human proteins and their variants

Giulia Babbi, Matteo Manfredi, Elisa Bertolini, Castrense Savojardo, Pier Luigi Martelli, Rita Casadio

Biocomputing Group, University of Bologna

Correspondence to: Rita Casadio, rita.casadio@unibo.it

Published: 6 Mar 2024

Abstract

Next-generation sequencing techniques provide an unprecedented characterisation of human Variants of Unknown Significance (VUS). Single-residue variations are collected in public databases and associated to diseases and phenotypes. However, for detailing at molecular level mechanisms involved in the onset of diseases, variants need structural and functional annotation. Here we propose a new portal called **Bioinformatic Sweeties**, collecting resources ranging from databases for human protein annotation to computational methods for predicting impact of variants. The tools, included in the portal, allow computing different protein properties, ranging from solvent accessible surface to stability and interactions and do not require login or installation. The portal, speeding up the variant characterisation process, is available at: https://bioinformaticsweeties.biocomp.unibo.it

Keywords: protein annotation, functional annotation, variant annotation, predictors, diseases

Introduction

Next-generation sequencing techniques constantly provide an unprecedented volume of human protein variants, whose majority is still of unknown significance (VUS). Single Nucleotide

Polymorphisms (SNPs) translate into protein variants, and in public databases they are associated to diseases and phenotypes. Following the classification of the American College of Medical Genetics and Genomics/Association for Molecular Pathology (ACMG/AMP) terminology[1], five kinds of protein variants are possible: i) pathogenic (P), ii) likely pathogenic (LP), iii) benign (B), iv) likely benign (LB), and v) of uncertain significance (US). These types are often grouped into three classes i) likely pathogenic/pathogenic (LP/P), ii) likely benign or benign (LB/B) and iii) of uncertain significance (US).

Considering the UniProt/Swiss-Prot[2] data base (https://www.uniprot.org/), the number of human protein entries associated with at least one variant is 13,058, covering 13,035 genes, and associated to a total of 82,499 variants. A recent statistics (Uniprot release 2024_01) lists 32,665 human variants, labelled LP/P, and related to 3,341 genes; 39,656 human variants labelled LB/B and related to 11,663 genes; and 10,178 US human variants related to 2,801 genes. Over the 20,428 human proteins in SwissProt, only 16% of the genes are disease related, for a total of 5,054 diseases.

In the last decade, different repositories for gene-disease and variant-disease associations, trying to be more comprehensive in including all studied gene variants, were implemented. ClinVar[3] (https://www.ncbi.nlm.nih.gov/clinvar/) and DisGeNet[4] (https://www.disgenet.org/) are among the largest databases collecting variants and their association to specific diseases, when known.

ClinVar (Current total, Jan 07-2024) lists 2,435,224 unique variation records, including variants on different protein isoforms/transcripts and many variation types, like Deletion, Duplication, Indel, Insertion and Single nucleotide. Among them, 1,952,891 (80%) have been deposited just by one single submitter or have conflicting interpretations. When constraining the list to LP/P human variants with gene sequence length <1kb, assessed by expert panels or multiple submitters, with a molecular consequence in the protein (Frameshift (16,748), Missense (13,662), Nonsense (14,242), Splice site (7,251)), the total number of protein variants is 51,615. These variants are associated with 2,445 genes.

DisGeNet, in turn, is a catalogue that aims to integrate data from expert-curated repositories, GWAS experiments, animal models and the scientific literature via text mining. DisGeNet lists 194,515 variants related to 14,155 diseases (including disorders, traits, and phenotypes). It also directly relates genes to diseases, connecting 21,671 genes to 30,170 diseases. Resources connecting genes to diseases include OMIM (https://www.omim.org/), one of the most curated resources of genetic diseases, listing 4,873 genes associated with 7,466 phenotypes, genetic disorders, and traits.

Focusing on genetic diseases, we can relate some 5,058 (25%) of the human proteins with 7,881 disease (see Figure 1).

Here we propose a new portal, **Bioinformatic Sweeties**, collecting resources from our group, ranging from databases for protein annotation to predictors for computing different structural/functional properties of variants. Our computational tools, state-of-the-art in the different tasks and recently published, may help in better annotating a protein in terms of disease/phenotypes associations.

In the portal we include four data bases whose major characteristics are the relation of the protein variants to biological pathways/processes, for a systemic comprehension of which processes may be hampered by the variations. A recent specific focus is also on human multifaceted proteins, including moonlighting and multifunctional human proteins. We include six recent predictors as well, to assess the impact of a variant on protein stability, to help in locating protein-protein interfaces, and to help relating variants to their effects on disease insurgence.

Results

Bioinformatic Sweeties unifies in a comprehensive and user-friendly portal (Figure 2) databases and web servers.

The home page of the portal (https://bioinformaticsweeties.biocomp.unibo.it) includes all the resources with a short description. A search bar is present, to help the user in finding the resources, along with a tag system to collect similar tools and quickly investigate the results filtering by a tag of interest.

A description of the tag-resource association and a general statistic of tag distribution is presented in Table 1.

Predictors:

ISPRED4

ISPRED4 (Interaction Site PREDictor, version 4) is a web-server[5] for predicting protein-protein interaction sites starting from protein structure. It adopts machine-learning methods (SVM+CRF) to predict interaction state of each residue in the protein surface by extracting several features from protein sequence and structure.

The server accepts in input a single protein structure, and the result page consists of two sections:

i) Sequence and Structure view:

this section includes general information about the protein (e.g. PDB ID, the protein length, the surface length). Residues at the surface or predicted as interfaces are highlighted in blue and red, respectively, both in a sequence view panel and in an interactive 3D-view window.

ii) Detailed report:

The table lists the results for each residue, detailing the Accessible Surface Area, the Relative Solvent Accessibility (RSA), the Predicted RSA, geometrical indexes (depth, protrusion, surface), the prediction of ISPRED4 and the associated probability. It also contains buttons to download the results in a text format or PDB format.

ISPRED-SEQ

ISPRED-SEQ [6] is a deep-learning based method for the prediction of Interaction Sites starting from protein sequence. The input sequence is firstly embedded using ProtTrans T5 and ESM-1v in order to produce a 2304-dimensional vectorial representation of each residue. This entirely substitutes the need for traditionally hand-crafted features such as physicochemical properties of the residues or sequence profiles derived from multiple sequence alignments.

There are three sections in the result page:

i) Visualize Results: this section contains general info about the job, namely the ID, the date of submission, the protein ID, the protein length and the number of Predicted Interaction Sites.

ii) Sequence features view: a graphical representation of predicted Interaction Sites (IS) and Non Interaction Sites (N), highlighted in yellow and blue, respectively, along with the computed probability.

iii) Tabular results: a table with the results for each residue, detailing the prediction and the probability, with the possibility to filter by prediction or residue type, along with buttons to download the results.

E-pRSA

E-pRSA is a method for predicting the Relative Solvent Accessibility of residues in a protein chain without requiring previous knowledge of the 3-dimensional structure.

The target sequence is firstly processed by two different and complementary protein language models (PLMs), ProtT5 and ESM2, to generate a concatenated vector of 1280+1024=2304 features for each residue. The output consists of a single value between 0 and 1, representing the putative RSA of the residue. A threshold of 20% is also adopted to distinguish Buried and Exposed residues.

The result page consists of three main sections: job information, feature viewer and data tables.

Job information section reports general information about the job, including the Job ID, the date of submission and completion, the protein ID, the protein length, the counts and percentages of exposed vs buried predictions, and the count and percentage of predicted interaction sites obtained with ISPRED-SEQ (see section above).

In case of a Batch Job, the number of proteins and total residues submitted are also shown, together with a button to download the results in a tab-separeted format.

The feature viewer shows the results visualized with the neXtProt feature viewer, where the first line displays the residues of the sequence, while the second and third lines show the output of E-pRSA as a regression (the putative RSA) and binary classification (Exposed or Buried), respectively. The last line shows predictions of ISPRED-SEQ highlighting residues that are likely protein-protein interaction sites. The data table section contains detailed predictions that can be filtered and combined using a set of filters, or downloaded for further analyses.

E-SNPs&GO

E-SNPs&GO[7] is a machine-learning method for predicting the pathogenicity of human variations.

It is a fast and accurate method that, given an input protein sequence and a single residue variation, can predict whether the variation is related to diseases or not. The Pathogenicity class (pathogenic or benign) is provided along with a pathogenicity probability and a reliability index (an integer number in the range [0-10] where 0 and 10 correspond to the minimum and maximum confidence for the prediction).

INPS

INPS (Impact of Non-synonymous mutations on Protein Stability) is a web server[8] for predicting the impact of non-synonymous Single Nucleotide Polymorphisms (nsSNPs) on protein stability starting from protein sequence. INPS is based on SVM regression and it is trained to predict the folding free energy change upon single-point variations in protein sequences.

When tested in cross-validation on a non-redundant dataset, INPS performs similarly to the state-of-the-art methods that also consider information on protein structure.

The results page contains two main sections, similarly to INSP-3D: the sequence view and the detailed mutation report.

INPS-3D

INPS-3D (Impact of Non-synonymous mutations on Protein Stability) is a web server[9] for predicting the impact of non-synonymous Single Nucleotide Polymorphisms (nsSNPs) on protein stability starting from protein structure. With respect to INPS (see above), the set of INPS3D descriptors includes features derived from protein 3D structure. In particular, we considered two additional structure-based features: i) the estimated local energy difference (ED) between native and mutated protein structures, and ii) the Relative Solvent Accessibility (RSA) of the native residue as computed from the PDB file using the DSSP method[10,11] and then normalized.

Results are presented in two main sections: the sequence view and the detailed mutation report. The Sequence view reports general information about the protein, including PDB ID and the protein length. Residues involved in prediction are highlighted in blue in a sequence panel. Detailed mutation report contains the prediction of Stability change (DDG) in kcal/mol for each submitted mutation, and a button to download the results in text format.

Databases

DAR

DAR (Disease And Reactome database) is a database[12] mapping disease-associated enzyme into Reactome pathways.

We adopt Reactome to describe human biological processes, and by mapping disease-associated enzymes in the Reactome pathways, we establish a Reactome-disease association.

This allows a novel categorization of human monogenic and polygenic diseases based on Reactome pathways and reactions.

The resource helps in dissecting the complexity of the human genetic disease universe, highlighting all the possible links within diseases and Reactome pathways.

Users may browse inserting a specific query, having as possible inputs:

- UniProt accession number
- Gene name
- Disease (MONDO)
- Disease (OMIM)
- Disease (Orphanet)
- Disease name
- Reactome ID
- Reactome name

- Reactome Root name
- EC number (4 digits)
- EC number (first digit)

MultifacetedProtDB

MultifacetedProtDB[13] is an integrated and manually curated database providing a comprehensive collection of multifunctional UniProt/SwissProt human proteins.

It includes 1103 proteins, of which 812 are enzymes. Only 241 human proteins were already present in other datasets of multifunctional proteins (Moon DB, MoonProt, and MultitaskProtDB II, which when restricted to humans collect 47, 103, and 185 proteins, respectively). Other proteins are derived from recent literature. Enzymes with multiple EC codes are included.

MultifacetedProtDB increases by four times the number of multifunctional proteins reported in currently available resources. For each protein the search allows retrieving structural and functional features alongside with inclusion in biological pathways and disease associations when present.

In the search bar, the user may insert a specific query, having as input:

- Gene name
- UniProt accession number or protein name
- Disease (MONDO, OMIM, Orphanet, ICD-10 or disease name
- Phenotypes (HPO ID or name)
- Sucellular location, Cell or tissue of expression
- EC number, Reactome (ID or name)
- GO term (ID or name)
- Pfam Interpro (ID or name)

eDGAR

eDGAR is a database[14] of Disease-Gene Associations with annotated Relationships among genes.

eDGAR collects and organizes data on gene/disease associations as derived from OMIM, Humsavar and ClinVar. For each heterogeneous or polygenic disease, eDGAR provides information on the relationship among the proteins encoded by the involved genes, including transcription factors and protein-protein interactions.

For each disease-associated gene eDGAR provides information on its annotation. Only genes associated to diseases are currently present in eDGAR.

There are many ways to enter eDGAR:Genes can be searched by HGNC code or Ensembl code (ENSG); Proteins can be searched by UniProt accession or by Ensembl code (ESNP); Diseases can be searched by OMIM code, including the phenotypic series code, or via text search.

eDGAR allows searching by a group of genes to retrieve the list of shared annotations.

PhenPath

PhenPath[15] is a web server for associating phenotypes with molecular functional annotations.

PhenPath includes a database and a tool: i) PhenPathDB collects all the functional annotations associated with a specific phenotype, and ii) PhenPathTOOL retrieves the annotations shared by two or more phenotypes; it can be used to retrieve diseases associated with a list of phenotypes, or to highlight the functional annotations enriched for the genes associated with a set of phenotypes.

The user can search PhenPathDB either starting from the general table (collecting OMIM Clinical Synopsis classification and HPO Phenotypic Abnormality subclassification), or specifying a query in the search page in terms of HPO (ID or name), Disease (OMIM ID or name), or OMIM clinical synopsis term.

Use case: PCBD1 and hyperphenylalaninemia variants

In this section we provide an example of use of the resources collected in Bioinformatics Sweeties for characterizing the possible effects of protein variants on pathogenic conditions. Fe focus on the enzyme pterin-4-alpha-carbinolamine dehydratase, encoded by the gene PCBD1 and involved in tetrahydrobiopterin biosynthesis.

We start our analysis searching the protein in DAR, with the corresponding UniProt accession "P61457". Results reported in Figure 3 show that the protein in an enzyme associated with a single EC number (4.2.1.96), a single disease (MONDO: 0009908, “pterin-4 alpha-carbinolamine dehydratase 1 deficiency”, a benign form of hyperphenylalaninemia due to tetrahydrobiopterin deficiency), and a single Reactome (R-HAS-8964208, “Phenylalanine metabolism”) with one reaction. In the interactive interface of DAR, the names associated to the codes can be retrieved by hovering over the links with the mouse.

To retrieve more information about the association with diseases, we search eDGAR using the same UniProt accession (Figure 4) and we found again the association with “hyperphenylalaninemia, BH4-deficient, D” (HPABH4D), having OMIM ID "264070".

It is autosomal recessive disorder characterized by mild transient hyperphenylalaninemia often detected by newborn screening, with increased excretion of 7-biopterin. Patients are almost asymptomatic, although infantile transient neurologic deficits may happen[16].Patients may also develop hypomagnesemia and nonautoimmune diabetes mellitus during puberty[17]. Among the others, the result page (Figure 4) contains the link to the UniProt webpage, with all protein variants, and the PDB identifiers and links to the available structure.

KEGG and Reactome pathways, GO terms, Transcription Factors and gene annotation as cytogenetic band or tandem repeats are reported, if available.

Searching in MultifacetedProtDB shows the protein is multifunctional (Figure 7) since it both prevents the formation of 7-pterins and accelerate the formation of quinonoid-BH2. It has also been proposed that this protein work as a coactivator for HNF1A-dependent transcription and in the dimerization of homeodomain protein HNF1A, enhancing its transcriptional activity, and it acts as a coactivator for HNF1B-dependent transcription[17].

The list of all the possible annotation fields found when searching MultifatedProtDB for gene PCBD1 is reported in Figure 5.

The Variant section of MutlifactedProtDb links the UniProt variant viewer and can be used to collect the pathogenic variants associated to the entry. In this use case, we focus on missense variants associated with the “hyperphenylalaninemia, BH4-deficient, D” (HPABH4D): T79I, C82R, R88Q, E97K.

Finally we submitted the 4 variants to E-SNPs&GO, that confirms their pathogenicity as reported in the literature (figure 6).

Since the structure of the protein has not been experimentally resolved, we use E-pRSA to predict the Relative Solvent Accessibility (RSA) for the positions of interest.

Three out of four positions are predicted to be exposed and in interaction sites (Figure 7), while position 79 is predicted to be buried.

We may then investigate the impact of the variants on protein stability using the INPS-SEQ (figure 8). It results that all the variants are predicted to promote a slight destabilization the protein (negative ΔΔG). The effect is more evident for variant T79I (ΔΔG -1.01 kcal/mol).

These results enable to formulate hypotheses on the mechanism of pathogenicity of the different variants: the only variant predicted to be buried has a significant impact on the protein stability T79I, while the other 3 variants (C82R, R88Q, E97K) are predicted to be only slightly destabilizing. In turn, all the three are predicted to be part of a protein-protein interfaces and then may alter protein function or pathways through modification of protein-protein interactions. Indeed, BioGRID[18], IntAct[19] and UniProt[2] report 90, 91 and 14 interactors, respectively.

Conclusions

Bioinformatic Sweeties is a portal collecting recently published resources for protein annotation and state-of-the-art predictors for structural/functional properties of protein variants. The usage of our computational tools may help in better annotating a protein and its variants in term of disease/phenotypes relations, also without an experimentally validated 3D structure, as discussed through two use cases.

Declarations

Funding

GB: Project title: "National Center for HPC, Big Data and Quantum Computing", code: CN00000013, CUP: J33C22001170001. Funded by the European Union - NextGenerationEU, PNRR - Mission 4 - Component 2 - Investment 1.4 "Strengthening research structures and creation of "national R&D champions" on some Key Enabling Technologies" D.D. 3138 of 12/16/2021 corrected with D.D. 3175 of 12/18/2021.

Conflict of Interest Declaration

The authors declare no conflict of interest.

Figure 1

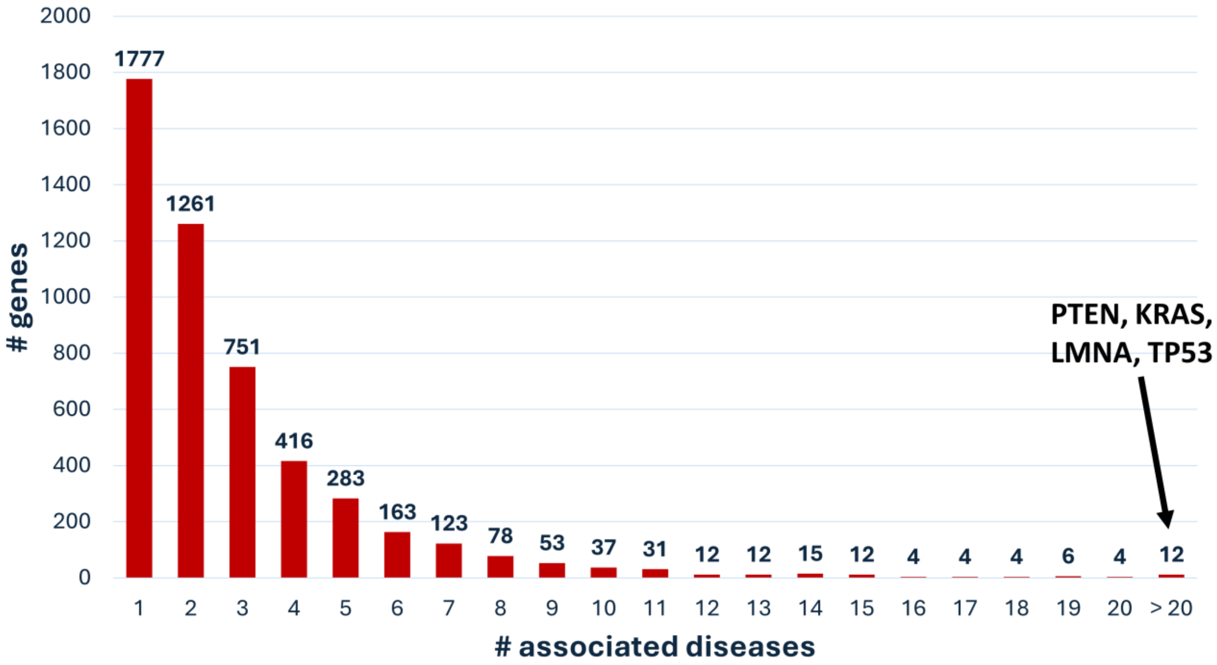
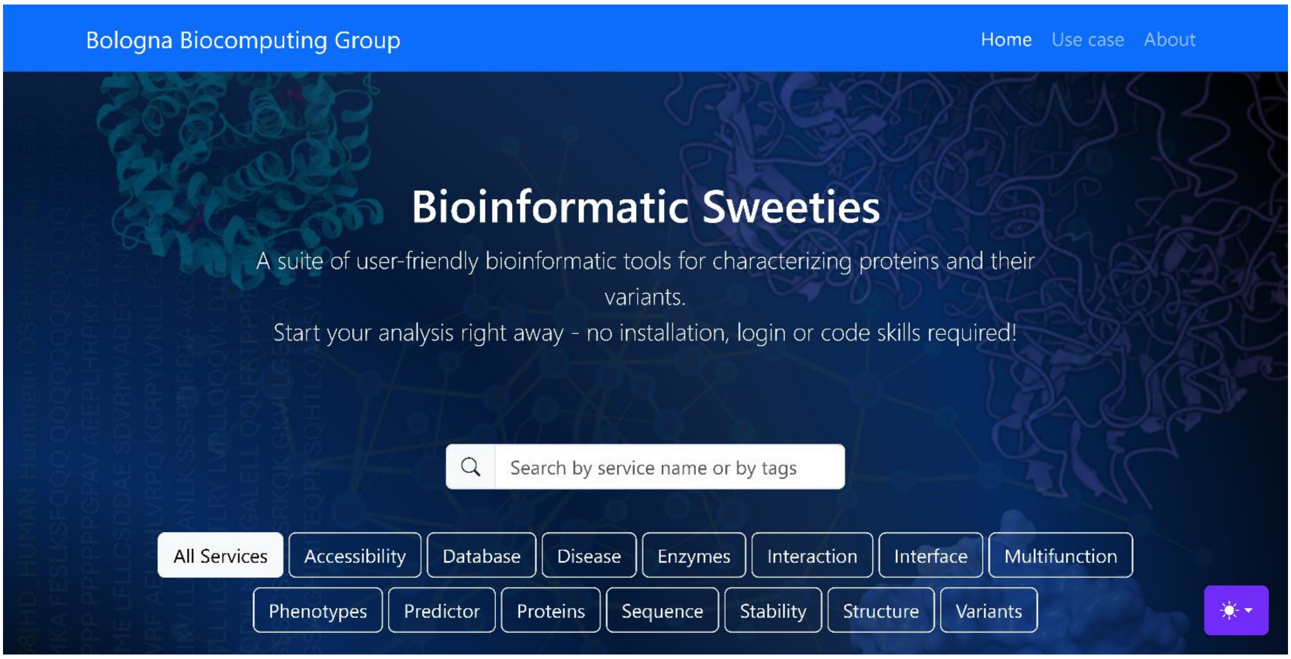


Figure 1: Distribution of the number of genes as a function of the number of associated diseases. The genes with the highest number of associated diseases are TP53 (encoding for the cellular tumor antigen p53 and associated with 34 diseases), LMNA (encoding for Prelamin-A/C and associated with 32 diseases), KRAS (encoding for the KRAS GTPase and associated with 30 diseases), and PTEN (endodng for the Inositol polyphosphate 3-phosphatase and associated with 28 diseases).

Figure 2



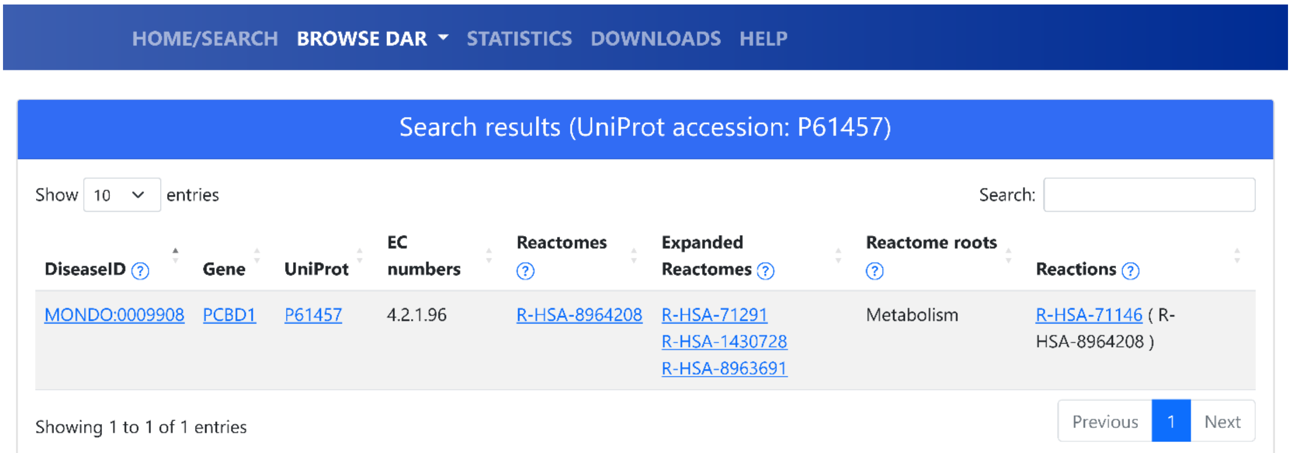
Bioinformatic Sweeties

A suite of user-friendly bioinformatic tools for characterizing proteins and their variants.

Start your analysis right away - no installation, login or code skills required!

Figure 2: Bioinformatic Sweeties Homepage (https://bioinformaticsweeties.biocomp.unibo.it)

Figure 3



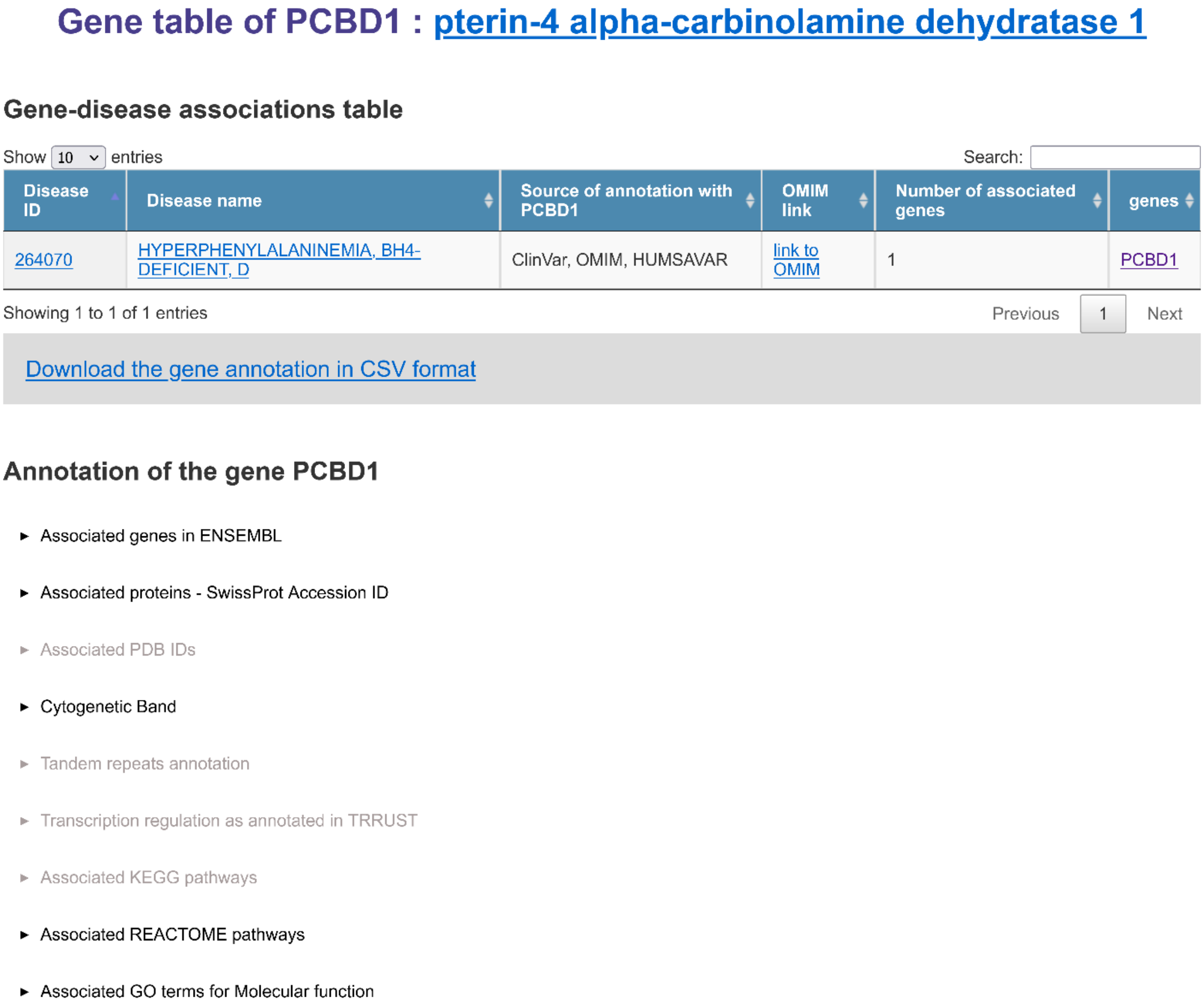
Search results (UniProt accession: P61457)

DiseaseID ?	Gene	UniProt	EC numbers	Reactomes ?	Expanded Reactomes ?	Reactome roots ?	Reactions ?
MONDO:0009908	PCBD1	P61457	4.2.1.96	R-HSA-8964208	R-HSA-71291 R-HSA-1430728 R-HSA-8963691	Metabolism	R-HSA-71146 (R-HSA-8964208)

Showing 1 to 1 of 1 entries

Figure 3: Details of the "Search Results" table in a DAR search.

Figure 4



Gene table of PCBD1 : pterin-4 alpha-carbinolamine dehydratase 1

Gene-disease associations table

Disease ID	Disease name	Source of annotation with PCBD1	OMIM link	Number of associated genes	genes
264070	HYPERPHENYLALANINEMIA, BH4-DEFICIENT, D	ClinVar, OMIM, HUMSAVAR	link to OMIM	1	PCBD1

Showing 1 to 1 of 1 entries

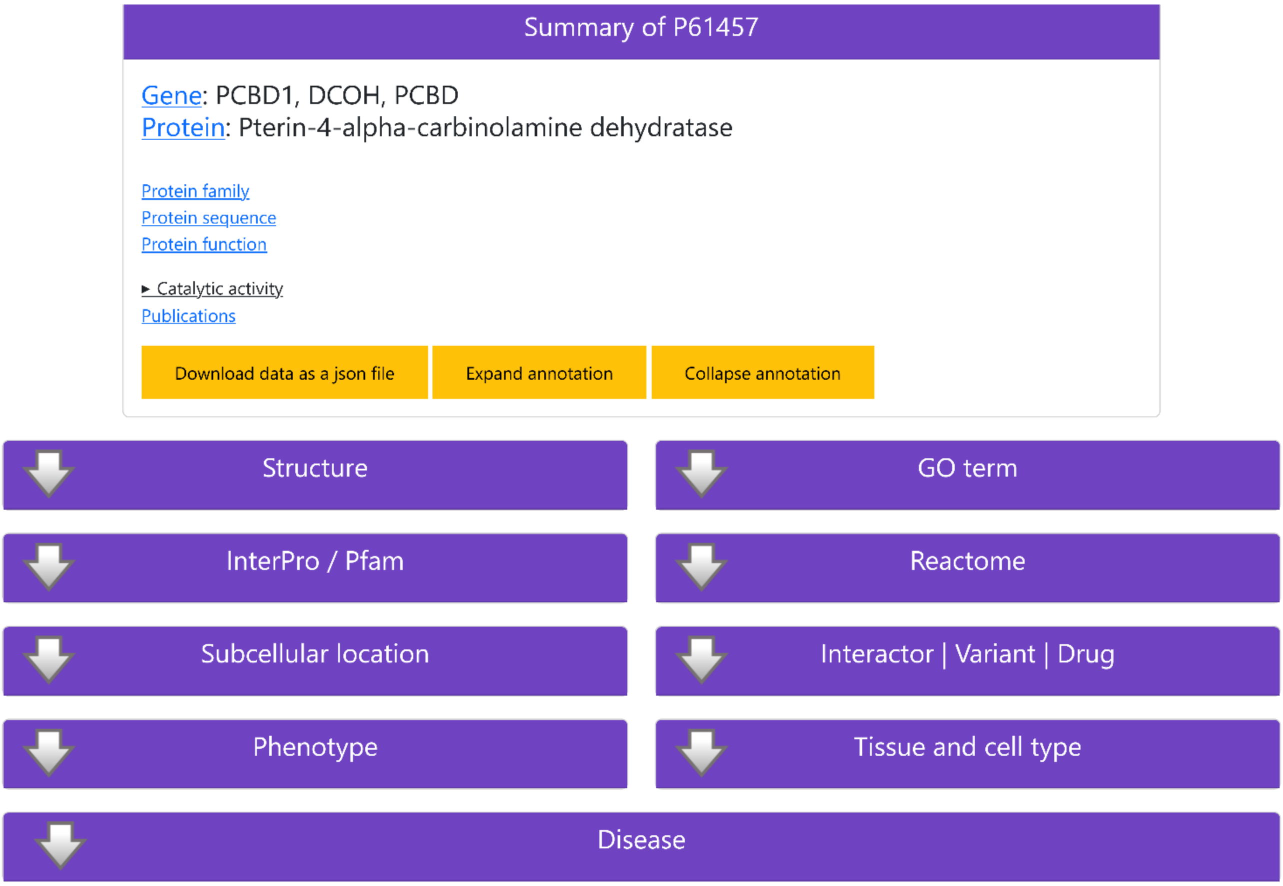
Download the gene annotation in CSV format

Annotation of the gene PCBD1

- ▸ Associated genes in ENSEMBL
- ▸ Associated proteins - SwissProt Accession ID
- ▸ Associated PDB IDs
- ▸ Cytogenetic Band
- ▸ Tandem repeats annotation
- ▸ Transcription regulation as annotated in TRRUST
- ▸ Associated KEGG pathways
- ▸ Associated REACTOME pathways
- ▸ Associated GO terms for Molecular function

Figure 4: Details of the "Gene-disease association table" and "Annotation of the gene" in a eDGAR search

Figure 5

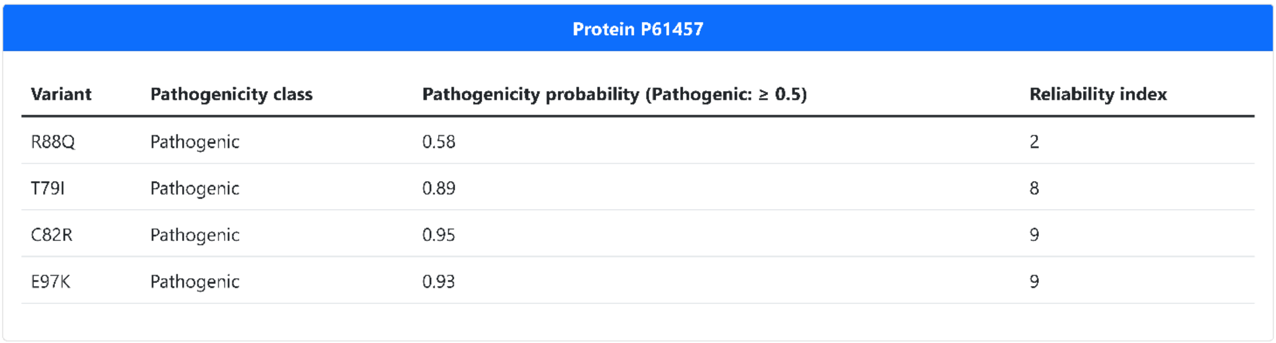


Summary of P61457

Gene: PCBD1, DCOH, PCBD
Protein: Pterin-4-alpha-carbinolamine dehydratase

Figure 5: Details of the result search in MultifactedProtDB.

Figure 6



Protein P61457

Variant	Pathogenicity class	Pathogenicity probability (Pathogenic: ≥ 0.5)	Reliability index
R88Q	Pathogenic	0.58	2
T79I	Pathogenic	0.89	8
C82R	Pathogenic	0.95	9
E97K	Pathogenic	0.93	9

Figure 6: Details of the results table in E-SNPs&GO.

Figure 7

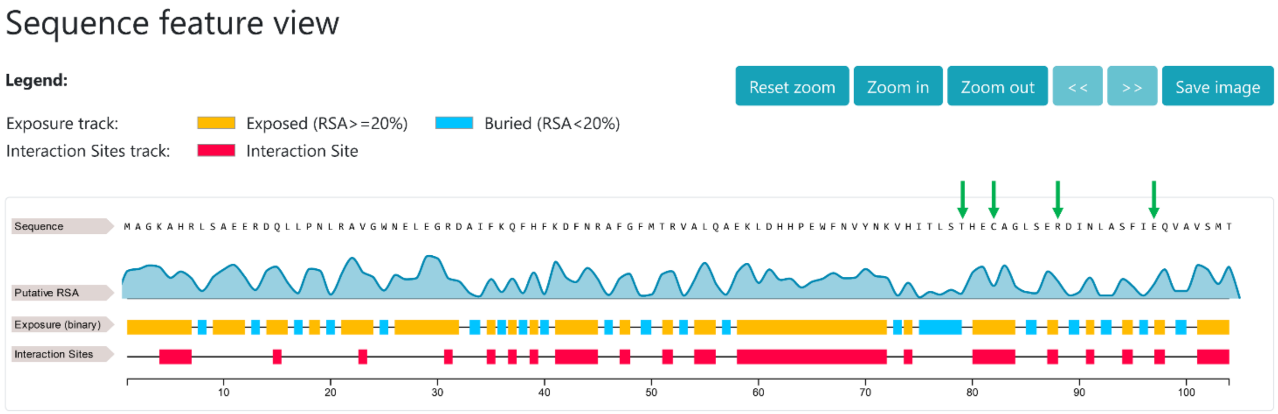
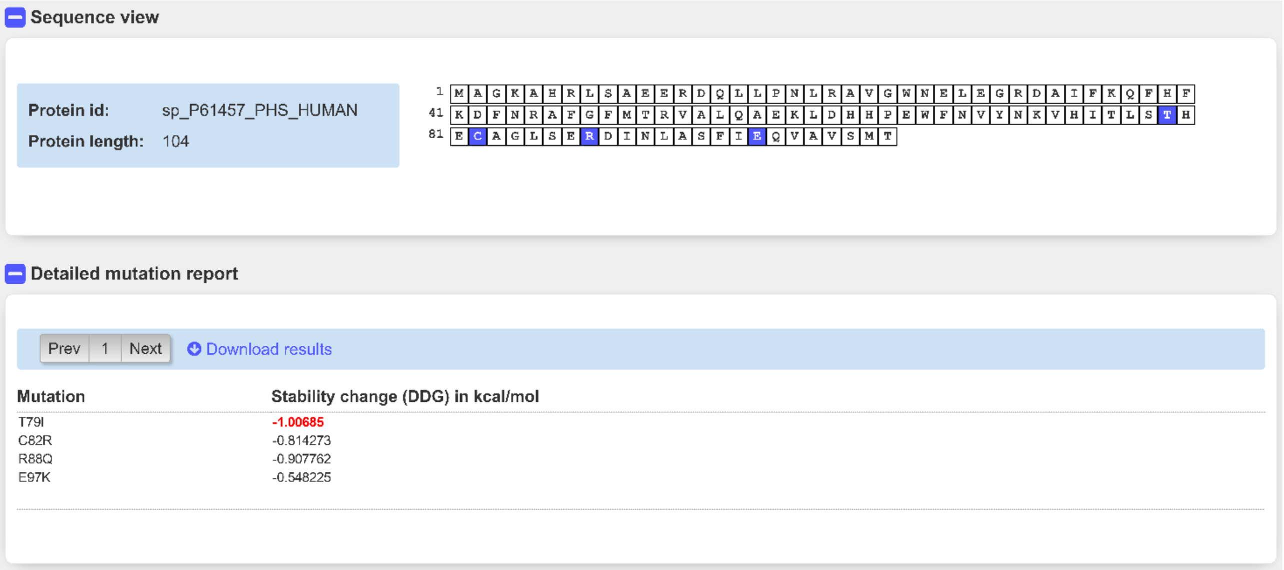


Figure 7: Details of the results of E-pRSA.

Figure 8


Sequence view

Protein id: sp_P61457_PHS_HUMAN

Protein length: 104

```
 1 M A G K A H R L S A E E R D Q L L P N L R A V G W N E L E G R D A I F K Q F H F
41 K D F N R A F G F M T R V A L Q A E K L D H H P E W F N V Y N K V H I T L S T H
81 E C A G L S E R D I N L A S F I E Q V A V S M T
```

Detailed mutation report

Mutation	Stability change (DDG) in kcal/mol
T79I	**-1.00685**
C82R	-0.814273
R88Q	-0.907762
E97K	-0.548225

Figure 8: Details of the prediction in INPS

Table 1

Tag	*# Resources*	*Resources*
Accessibility	1	E-pRSA
Database	4	DAR, MultifacetedProtDB, PhenPath, eDGAR
Disease	2	DAR, E-SNPs&GO, eDGAR, PhenPath
Enzymes	1	DAR
Interaction	2	ISPRED-SEQ, ISPRED4
Interface	2	ISPRED-SEQ, ISPRED4
Multifunction	1	MultifacetedProtDB
Phenotypes	1	PhenPath
Predictor	6	E-SNPs&GO, E-pRSA,INPS, INPS-3D, ISPRED-SEQ, ISPRED4
Proteins	3	E-pRSA, MultifacetedProtDB, eDGAR
Sequence	4	E-SNPs&GO, E-pRSA, INPS, ISPRED-SEQ
Stability	2	INPS, INPS-3D
Structure	2	INPS-3D, ISPRED4
Variants	3	E-SNPs&GO, INPS, INPS-3D

Table 1: Tags in Bioinformatic Sweeties and their associated resources.

References

1. Richards, S. *et al.* 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. Off. J. Am. Coll. Med. Genet.* **17**, 405–424 (2015).

2. UniProt: the Universal Protein Knowledgebase in 2023 | Nucleic Acids Research | Oxford Academic. https://academic.oup.com/nar/article/51/D1/D523/6835362?login=true.

3. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).

4. DisGeNET knowledge platform for disease genomics: 2019 update | Nucleic Acids Research | Oxford Academic. https://academic.oup.com/nar/article/48/D1/D845/5611674?login=true.

5. ISPRED4: interaction sites PREDiction in protein structures with a refining grammar model | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/33/11/1656/2953248.

6. Manfredi, M., Savojardo, C., Martelli, P. L. & Casadio, R. ISPRED-SEQ: Deep Neural Networks and Embeddings for Predicting Interaction Sites in Protein Sequences. *J. Mol. Biol.* **435**, 167963 (2023).

7. Manfredi, M., Savojardo, C., Martelli, P. L. & Casadio, R. E-SNPs&GO: embedding of protein sequence and function improves the annotation of human pathogenic variants. *Bioinformatics* **38**, 5168–5174 (2022).

8. INPS: predicting the impact of non-synonymous variations on protein stability from sequence | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/31/17/2816/183893.

9. INPS-MD: a web server to predict stability of protein variants from sequence and structure | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/32/16/2542/1743481.

10. Touw, W. G. *et al.* A series of PDB-related databanks for everyday needs. *Nucleic Acids Res.* **43**, D364–D368 (2015).

11. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).

12. Mapping human disease-associated enzymes into Reactome allows characterization of disease groups and their interactions | Scientific Reports. https://www.nature.com/articles/s41598-022-22818-5.

13. MultifacetedProtDB: a database of human proteins with multiple functions | Nucleic Acids Research | Oxford Academic. https://academic.oup.com/nar/article/52/D1/D494/7288824.

14. Babbi, G. *et al.* eDGAR: a database of Disease-Gene Associations with annotated Relationships among genes. *BMC Genomics* **18**, 554 (2017).

15. Babbi, G., Martelli, P. L. & Casadio, R. PhenPath: a tool for characterizing biological functions underlying different phenotypes. *BMC Genomics* **20**, 548 (2019).

16. Thöny, B. *et al.* Hyperphenylalaninemia with high levels of 7-biopterin is associated with mutations in the PCBD gene encoding the bifunctional protein pterin-4a-carbinolamine dehydratase and transcriptional coactivator (DCoH). *Am. J. Hum. Genet.* **62**, 1302–1311 (1998).

17. Ferrè, S. *et al.* Mutations in PCBD1 Cause Hypomagnesemia and Renal Magnesium Wasting. *J. Am. Soc. Nephrol.* **25**, 574 (2014).

18. Oughtred, R. *et al.* The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci.* **30**, 187–200 (2021).

19. Orchard, S. *et al.* The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42**, D358–D363 (2014).