TargetSNPdb: a database of preliminary analysis data of the impact of nsSNPs on drug target, drug metabolizing enzyme and disease associated genes

S.A. de Alencar[1,2*], E.C. Santos[1,2] , A.M. José[2] and J.C.D. Lopes[2]

[1]Departamento de Bioquímica e Imunologia, Bioinformática, Universidade Federal de Minas Gerais, Av. Antônio Carlos 6627, Belo Horizonte – M.G., 31270-901, Brazil

[2]Chemoinformatics Group, NEQUIM, Departamento de Química, Universidade Federal de Minas Gerais, Av. Antônio Carlos 6627, Belo Horizonte – M.G., 31270-901, Brazil

ABSTRACT

Motivation: The presence of nsSNPs in genes encoding drug targets, or drug metabolizing enzymes has been increasingly associated with drug response and diseases. The use of computational tools to analyze sequence and structure data of proteins can contribute to increase prediction efficiency of the impact caused by these nsSNPs.

Results: We have developed TargetSNPdb, a database server that contains computational predictions of the structural and functional impact of nsSNPs in protein coding genes, including drug target and drug metabolizing enzyme encoding genes. The analysis results obtained from several computational tools (such as SIFT, PolyPhen, AutoDock, and GROMACS) relevant to the study of the impact of amino acid residue substitutions were integrated to existent information records from the literature and genetic association databases, enabling the combination of results from a variety of different approaches to evaluate the potential impact of nsSNPs on protein function. Potential applications of TargetSNPdb include the priorization of nsSNPs for association and experimental studies.

Availability: The TargetSNPdb can be accessed via http://nequim.qui.ufmg.br/targetsnp/

Contact: sergiodealencar@gmail.com

INTRODUCTION

Single nucleotide polymorphisms (SNPs) constitute the most frequent type of sequence variation in humans, making up about 90% of all human genetic variation. Currently, there are over 30 million human SNPs listed in publicly accessible databases, of which over 210,000 are located within protein coding sequences (http://www.ncbi.nlm.nih.gov/SNP/). A fraction of these coding SNPs which alter the encoded amino acid sequence are known as non-synonymous SNPs (nsSNPs) (Sachidanandam et al., 2001).

The presence of nsSNPs in genes coding drug targets, or drug metabolizing enzymes, can cause structural variations in the active site of these proteins and, as a result, could affect drug interaction or destabilize the complex formed (Rajasekaran et al., 2008). Also, changes in stability, which could be caused by a reduction in hydrophobic area, overpacking, backbone strain, or loss of electrostatic interactions, may affect a protein's folding rate and increase its susceptibility to proteolysis, resulting in reduced

https://doi.org/10.26434/chemrxiv-2024-8nhvq **ORCID:** https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

concentration of the native protein, and diseases (Wang et al., 2001; Yue et al., 2005; Karchin et al., 2005). Therefore, nsSNPs are critical to understand the efficiency and toxicity of drugs.

The use of Bioinformatics and Chemoinformatics computational tools to analyze available sequence and structure data of proteins can contribute to increase prediction efficiency of the impact caused by nsSNPs on protein coding genes (Kapetanovic, 2008). Several studies have shown that the impact caused by the substitution of amino acid residues on protein structures can be predicted by using both a sequence homology based tool (SIFT) and a structural homology based method (PolyPhen) (Rajasekaran et al., 2007; Doss et al., 2008; Doss et al., 2008b; Rajasekaran et al., 2009), and that molecular docking can be useful in predicting possible changes in ligand interaction energies between native and variant drug targets (Purohit et al., 2008).

Hence, the rapid accumulation of new data of human nsSNPs and drug target (and metabolizing enzyme) protein sequence and structure, together with computational analysis results, is opening the way to improve understanding of the relationships between genotype, drug response, and disease. However, at present, relevant nsSNP and protein target information are scattered across many databases, and the computational prediction of the impact of nsSNPs on drug targets is limited to a few receptors (Bigler et al., 2007; Liu et al., 2009), creating new challenges for linking genetic variation with drug response variation.

We propose a database to collect, analyze and integrate as much as possible of the molecular level data relevant to the mechanisms that link nsSNP records to drug related information. TargetSNPdb is a Bioinformatics database that describes nsSNP records data, frequency information, nsSNP prediction of potential impact results, molecular docking and stability comparisons between native and mutant structures, association studies from the literature, and mapping of nsSNP positions in drug target and drug metabolizing enzyme structures.

METHODS

Database setup

TargetSNPdb was implemented in MySQL, version 5.1.45 (http://www.mysql.com/), a freely available relational database management system (RDBMS), and its graphical CGI interface was programmed in PHP, version 5.2.8 (http://php.net), using ADOdb, version 5.11 (http://adodb.sourceforge.net), an open source database abstraction library for PHP. The software DBDesigner, version 4.0.5.6 (http://www.fabforce.net/dbdesigner4) was used to model the data (Figure 1). The database is maintained on a DELL PowerEdge server using Ubuntu Linux, version 10.04.

Contents of TargetSNPdb

nsSNP data

Information about human nsSNP records was obtained from dbSNP build 131 (dbSNP Build:131), a resource at the National Center of Biotechnology Information that catalogs SNPs (Sherry et al., 2001). The following limits were used: Organism (Homo sapiens), Function Class (coding non- synonymous missense), and SNP Class (SNP). All redundant nsSNP records which have been merged to existent nsSNP records were removed. Population frequency data of nsSNP records was obtained from the International HapMap Project Biomart site (Thorisson et al., 2005) using the following parameters: Schema (rel22_NCBI_Build36), Database (HapMap_rel22), Dataset (All Populations), and filtering only nsSNPs and alleles with a frequency [>=] 0.01.

https://doi.org/10.26434/chemrxiv-2024-8nhvq **ORCID:** https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Prediction of the impact of nsSNPs on protein function

The SIFT algorithm predicts whether an amino acid substitution affects protein function based on sequence homology among related genes and domains over evolutionary time, and the physical-chemical properties of the amino acid residues (Ng and Henikoff, 2001; Ng and Henikoff, 2002; Ng and Henikoff, 2006). SIFT takes a query sequence and uses multiple alignment information to predict tolerated and deleterious substitutions for a position of interest in the query sequence (Ng and Henikoff, 2003). It is a multistep procedure that, given a protein sequence, (1) searches for similar sequences in a protein database, (2) chooses closely related sequences that may share similar function, (3) obtains the multiple alignment of these chosen sequences, and (4) calculates normalized probabilities for a chosen substitution in a given position in the alignment. Substitutions at each position with normalized probabilities less than a tolerance index of 0.05 are predicted to be intolerant or deleterious; those greater than or equal to 0.05 are predicted to be tolerated (Ng and Henikoff, 2001; Ng and Henikoff, 2006).

Prior to running SIFT locally, a protein database in FASTA format was obtained from EMBL (ftp://ftp.ebi.ac.uk/pub/databases/fastafiles/uniprot/) and properly parsed for recognition by the program. Using the relational database built, a selection was made of all nsSNP entries which could be mapped to SwissProt entries, and this data was used to build the input files (containing information about the amino acid residue substitutions and their respective protein sequence files) required to run SIFT. Sequence conservation and the nature of the amino acid residues involved in a substitution are also incorporated by PolyPhen-2, but it also values the location of the substitution within known structures and structural features of the protein available in the annotated database SwissProt (Adzhubei, 2010). For an amino acid residue substitution, PolyPhen-2 calculates Naïve Bayes posterior probability that this mutation is damaging and reports estimates of false positive and true positive rates. A mutation is also appraised qualitatively, as benign, possibly damaging, or probably damaging based on the model's false positive rate. If the lack of data does not allow to make a prediction, then the outcome is reported as unknown, and it has been represented in TargetSNPdb as N/A (Adzhubei, 2010). Publicly available pre-computed PolyPhen-2 annotations for dbSNP build 131 (http://genetics.bwh.harvard.edu/pph2/dokuwiki/downloads) was incorporated into TargetSNPdb.

Association of nsSNP records with diseases or literature records

Information about disease associated nsSNP records described in the Genetic Association Database (Becker et al., 2004), and nsSNPs records linked to PubMed entries and OMIM were included in TargetSNPdb (McKusick et al., 1998).

Protein data

Protein structural and sequence data along with annotations of function, pathway and family were obtained from the PDB, SwissProt and PANTHER databases (Bermanet al., 2000; Gasteiger et al., 2001; Thomas et al., 2003). Additional sequence and structure information about the location of variant amino acid residues in the SwissProt sequence was obtained from the SwissProt Variant Pages (Yip et al., 2004).

Drug related information

All information related to drugs (drug entries, drug targets, and drug metabolizing enzymes) was obtained from DrugBank (Wishart et al., 2008), KEGG (Kanehisa et al., 2010), and TTD databases (Zhu et al., 2009).

https://doi.org/10.26434/chemrxiv-2024-8nhvq **ORCID:** https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Protein Side Chain Modeling

We retrieved from the Protein Data Bank all native three-dimensional crystal structures available which were coded by genes which contained nsSNPs (Berman et al., 2000). Information about positions of the nsSNPs on the PDB native structures was obtained from the coliSNP database (Kono et al., 2008). Amino acid residue substitutions corresponding to nsSNPs in the native proteins were performed using the software SCWRL version 4, one of the most accurate programs used for protein side-chain modeling (Krivov et al., 2009).

Stability Analysis

In order to evaluate and compare the stability of native and modeled mutant structures generated with SCWRL4, energy minimization of the modeled 3D structures were done using the GROMACS software version 4.0 (Hess et al., 2008). The algorithms used for energy minimization was steepest descent (6000 steps). The stability change value was calculated as the Potential Energy change (in Kcal/mol) between the native and variant protein structures using the GROMOS G53a6 force field (Oostenbrink et al., 2004).

Molecular Docking Analysis

All ligands crystallized in complex with drug targets which were coded by genes containing nsSNPs were selected for docking studies. Molecular docking calculations were carried out using the public software AutoDock 4.0 (Morris et al., 2009). Before the docking process, grid maps representing the interaction energies between the various ligand atom types and the amino acid residue atoms in the receptor active site were calculated with the AutoGrid package of AutoDock. The center of the grid was defined as the center of the receptor active site, with points spaced at 0.375 Å.

Using the AutoDockTools (ADT) (Morris et al., 2009) package, polar hydrogen atoms were added geometrically to the protein structures, and partial atomic charges were calculated using the Gasteiger-Marsili method. ADT was also used to assign the number of torsions and to add polar hydrogen atoms to each of the ligand structures.

Docking experiments were done using the Lamarckian genetic algorithm for the global search, and the Solis-Wets algorithm for the subsequent local optimization. The actual population comprised 150 individuals. We set the maximum number of energy evaluations accordingly to the number of degrees of freedom of the ligands studied (ranging from 1-6 million energy evaluations), the maximum number of generations to 270,000 and the number of runs to 100. A maximal mutation rate of 0.02, an elitism of 1, a crossover rate of 0.8 and a local search rate of 0.06 were used. Default values were used for all remaining parameters.

A full description of all data types and number of entries that have been loaded into TargetSNPdb is shown in Table 1.

RESULTS AND DISCUSSION

TargetSNPdb can be accessed through a web-based interface, which was constructed using php scripts to communicate with the MySQL database. The interface was designed to offer a variety of searching options: SNP RS Number, Gene symbol, SwissProt AC, PDB Code, Protein Name, Drug Target Name, Drug Name, Metabolizing Enzyme Name, Pathway Name, and OMIM Phenotype Info (Figure 2A). Full list of drug target names, drug names, drug metabolizing enzyme names, pathway names, and OMIM Phenotyping Info names are also provided in the TargetSNPdb main web-page for facilitating the search of particular entries.

https://doi.org/10.26434/chemrxiv-2024-8nhvq **ORCID:** https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

The search is case insensitive, and incomplete form of names (or characters) can be used in all search fields. For instance, the input of “acetyl” finds entries with drug target name composed of characters “acetyl” such as “Acetylcholine” and “Acetyl-CoA carboxylase 2”. The wild character “%” can also be used in a search to allow for more flexibility. For example, the input of “tyrosine%kinase” in the drug target name search field finds entries whose drug target name contains both “tyrosine” and “kinase”, such as “Tyrosine-protein kinase Lck”. The character “%” here represents a string of arbitrary characters of any length.

The result of each search is displayed as a table, in which each column corresponds to information relevant to the search chosen by the user, such as a search by drug target name “dehydrogenase class 4 mu/sigma chain”(Figure 2B). In this example, all the drug target names that satisfy the search criteria are listed along with its SwissProt AC, SwissProt IC, SwissProt Variant ID, Variants, SNP RS Number and computational predictions. More detailed information about the variation contained in the protein can be obtained by clicking the corresponding SwissProt Variation ID or SNP RS Number. The result is displayed in another window, from which one may find information about the location of the variation in the protein structure, protein sequence, protein stability information, physical chemical properties, surface accessibility of the native and variant amino acid residues, and the computational prediction of the impact of the variation.

In our laboratory, TargetSNPdb is currently being used to search for associations between drug response and diseases. The advantage of combining scores and analysis results produced by different methods, such as SIFT, PolyPhen, optimization, and molecular docking, is that each method uses different algorithms, so that when the results obtained from all these agree, predictions are more trustworthy. Also, if nsSNPs are associated with known drug responses or diseases, these combined predictions might explain the association. Future developments include the integration of a database containing experimentally determined drug affinity data, and updates for newly released dbSNP builds.

Table 1. Data types and number of entries that have been loaded into TargetSNPdb.

Data Type	Number of entries
Literature Info	3024
SNP Frequency Info	56164
SIFT Analysis	88127
PolyPhen Analysis	69844
SNP Info	103207
Metabolization Info	163
Gene Info	45415
SwissProt Variation	56537
SNP PDB	5699
PDB Info	59330
Stability Analysis	4402
Docking Analysis	311
OMIM Info	1862
Drug Target Info	7265
Pathway Info	162

https://doi.org/10.26434/chemrxiv-2024-8nhvq **ORCID:** https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

ACKNOWLEDGEMENTS

Funding: This work was supported by a FAPEMIG fellowship (S.A. de Alencar, and E.C. Santos), a CNPQ fellowship (A.M. José), and a CNPQ grant (J.C.D. Lopes).

REFERENCES

Adzhubei,I.A. et al. (2010) A method and server for predicting damaging missense mutations. Nat Methods, 7(4):248–249.

Becker,K.G. et al. (2004). The genetic association database. Nat Genet., 36(5):431– 432.

Berman,H.M. et al. (2000). The protein data bank. Nucleic Acids Res., 28(1):235–242.

Gasteiger,E. et al. (2001). SWISS-PROT: connecting biomolecular knowledge via a protein database. Curr Issues Mol Biol., 3(3):47–55.

George Priya Doss,C. et al. (2008a). A novel computational and structural analysis of nsSNPs in CFTR gene. Genomic Medicine, 2(1-2):23–32.

George Priya Doss,C. et al. (2008b). Identification and structural comparison of deleterious mutations in nsSNPs of ABL1 gene in chronic myeloid leukemia: A bio-informatics study. J Biomed Inform., 41(4):607–612.

Hess,B. et al. (2008). GROMACS 4: Algorithms for highly efficient, Load-Balanced, and scalable molecular simulation. J Chem Theory Comput., 4(3):435–447.

Kanehisa,M. et al. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res., 38:D355–D360.

Kapetanovic,I.M. (2008). Computer-aided drug discovery and development (CADDD): in silico-chemico-biological approach. Chem Biol Interact., 171(2):165–176.

Karchin,R. et al. (2005). LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. Bioinformatics, 21(12):2814–2820.

Kono,H., et al. (2008). coliSNP database server mapping nsSNPs on protein structures. Nucleic Acids Res., 36:D409-13.

Krivov,G.G. et al. (2009). Improved prediction of protein side-chain conformations with SCWRL4. Proteins: Struc Func Bioinf., 77(4):778–795.

Morris,G.M. et al. (2009). AutoDock4 and AutoDockTools 4: Automated docking with selective receptor flexibility. J Comput Chem., 30(16):2785–2791.

Ng,P.C. and Henikoff,S. (2001). Predicting deleterious amino acid substitutions. Genome Res., 11(5):863–874.

Ng,P.C. and Henikoff,S. (2002). Accounting for human polymorphisms predicted to affect protein function. Genome Res., 12(3):436–446.

Ng,P.C. and Henikoff,S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res., 31(13):3812–3814.

Ng,P.C. and Henikoff,S. (2006). Predicting the effects of amino acid substitutions on protein function.

https://doi.org/10.26434/chemrxiv-2024-8nhvq **ORCID:** https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

Annu Rev Genomics Hum Genet., 7(1):61–80.

Oostenbrink,C. et al. (2004). A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. J Comp Chem., 25(13):1656–1676.

Rajasekaran,R. et al. (2008). Effect of deleterious nsSNP on the HER2 receptor based on stability and binding affinity with herceptin: a computational approach. C R Biol., 331(6):409–417.

Rajasekaran,R. and Sethumadhavan,R. (2010). In silico identification of significant detrimental missense mutations of EGFR and their effect with 4-Anilinoquinazoline-based drugs. Appl Biochem Biotechnol., 160(6):1723–1733.

Rajasekaran,R. et al. (2007). Identification and in silico analysis of functional SNPs of the BRCA1 gene. Genomics, 90(4):447–452.

Ramensky,V. et al. (2002). Human non-synonymous SNPs: server and survey. Nucleic Acids Res., 30(17):3894–3900.

Rituraj,P. et al. (2008). Studies on flexibility and binding affinity of asp25 of HIV-1 protease mutants. Int J Biol Macromol., 42(4):386-91.

Sachidanandam,R. et al. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature, 409(6822):928- 933.

Sherry,S.T. et al. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res., 29(1):308–311.

Sunyaev,S. (2001). Integration of genome data and protein structures: prediction of protein folds, protein interactions and 'molecular phenotypes' of single nucleotide polymorphisms. Curr Opin Struct Biol., 11(1):125–130.

Thomas,P.D. et al. (2003). PANTHER: a library of protein families and subfamilies indexed by function. Genome Res., 13(9):2129–2141.

Thorisson,G.A. et al. (2005). The international HapMap project web site. Genome Res., 15(11):1592–1593.

Wang,Z. and Moult,J. (2001). SNPs, protein structure, and disease. Hum Mutat., 17(4):263–270.

Wishart,D.S. et al.(2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res., 36:D901–906.

Yip,Y.L. et al. (2004). The Swiss-Prot variant page and the ModSNP database: A resource for sequence and structure information on human protein variants. Hum Mutat., 23(5):464–470.

Yue,P. et al. (2005). Loss of protein structure stability as a major causative factor in monogenic disease. J Mol Biol., 353(2):459–473.

Figure 1. Data model schema showing the relational structure of TargetSNPdb, and all the tables and their relationships. A line with an empty diamond represents a one-to-one relationship while a half- filled diamond represents a one-to-many relationship. Primary keys are indicated with a key.

https://doi.org/10.26434/chemrxiv-2024-8nhvq **ORCID:** https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

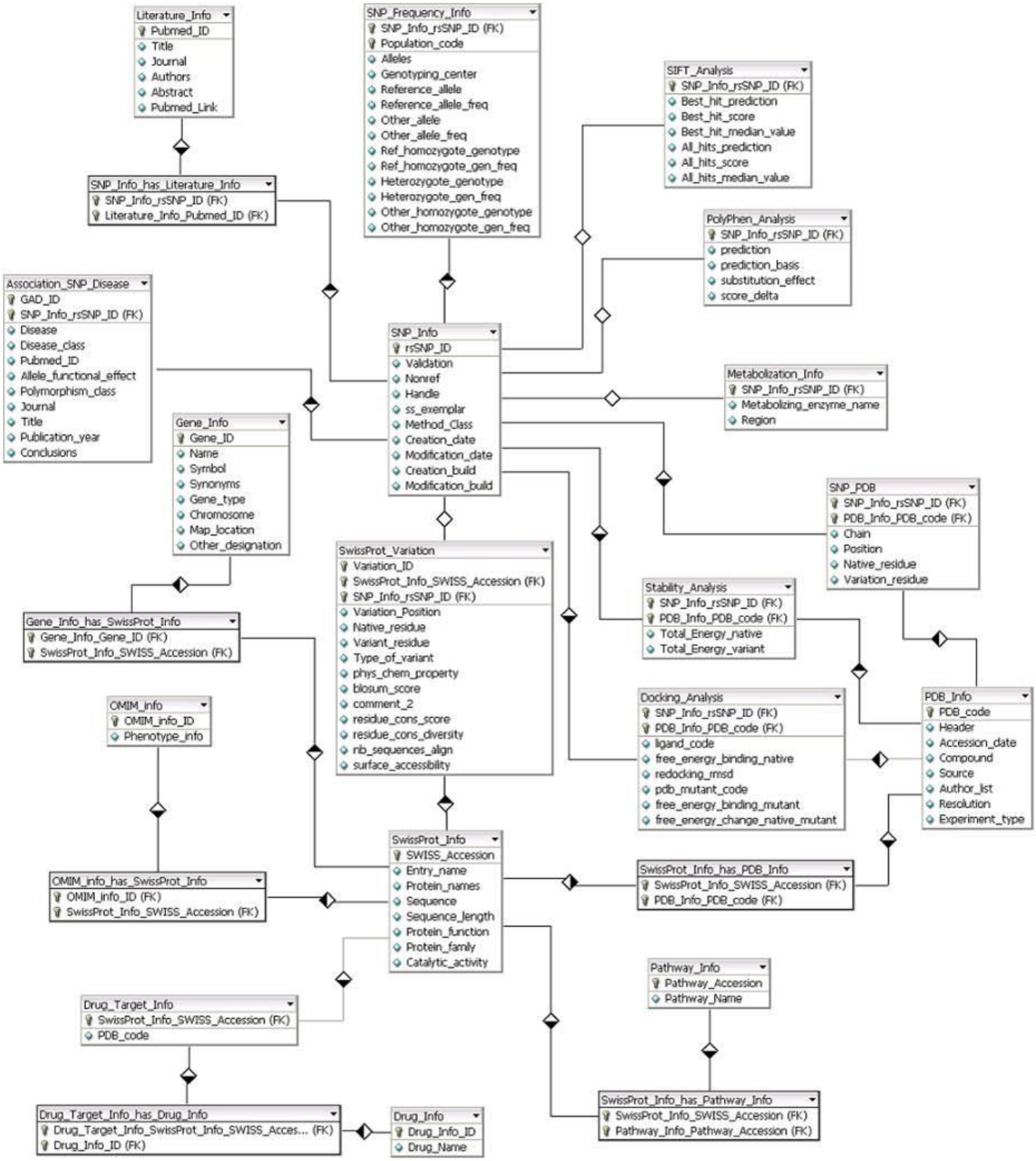


Figure 2. (A) A screenshot montage of the TargetSNPdb interface showing several possible search options available for the user. (B) Overview of a result returned by querying TargetSNPdb using the drug target name search option (selecting Alcohol dehydrogenase class 4 mu/sigma chain). The blue arrows point to the information contained in each hyperlink shown in the intermediate results page.

https://doi.org/10.26434/chemrxiv-2024-8nhvq **ORCID:** https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

A

SNP RS Number (e.g. 4531, 2020914):

OK

Gene symbol (e.g. CDH2, PCK1):

OK

SwissProt AC (e.g. P52630, O95477):

OK

PDB (e.g. 10GS, 1A22):

OK

Drug Target Name:

type above or select here
1-aminocyclopropane-1-carboxylate synthase-like protein 1 (ACC synthase-like protein)
1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase delta-3 (EC 3.1.4.11) (Phosphoinositide phospholipase C delt
10-formyltetrahydrofolate dehydrogenase (10-FTHFDH) (EC 1.5.1.6) (Aldehyde dehydrogenase family 1 member L1)

OK

Drug Metabolizing Enzyme Name:

type above or select here
adenosine_A2a_receptor
adrenergic_alpha_1A_receptor
adrenergic_beta-1_receptor

OK

Pathway Name:

type above or select here
2-arachidonoylglycerol biosynthesis
5-Hydroxytryptamine biosynthesis
5-Hydroxytryptamine degredation

OK

Drug:

Drug Name:

type above or select here
(+)-Irinotecan
(-)-3PPP, Maryland
(1'r,2's)-9-(2-Hydroxy-3'-Keto-Cyclopenten-1-Yl)Adenine

OK

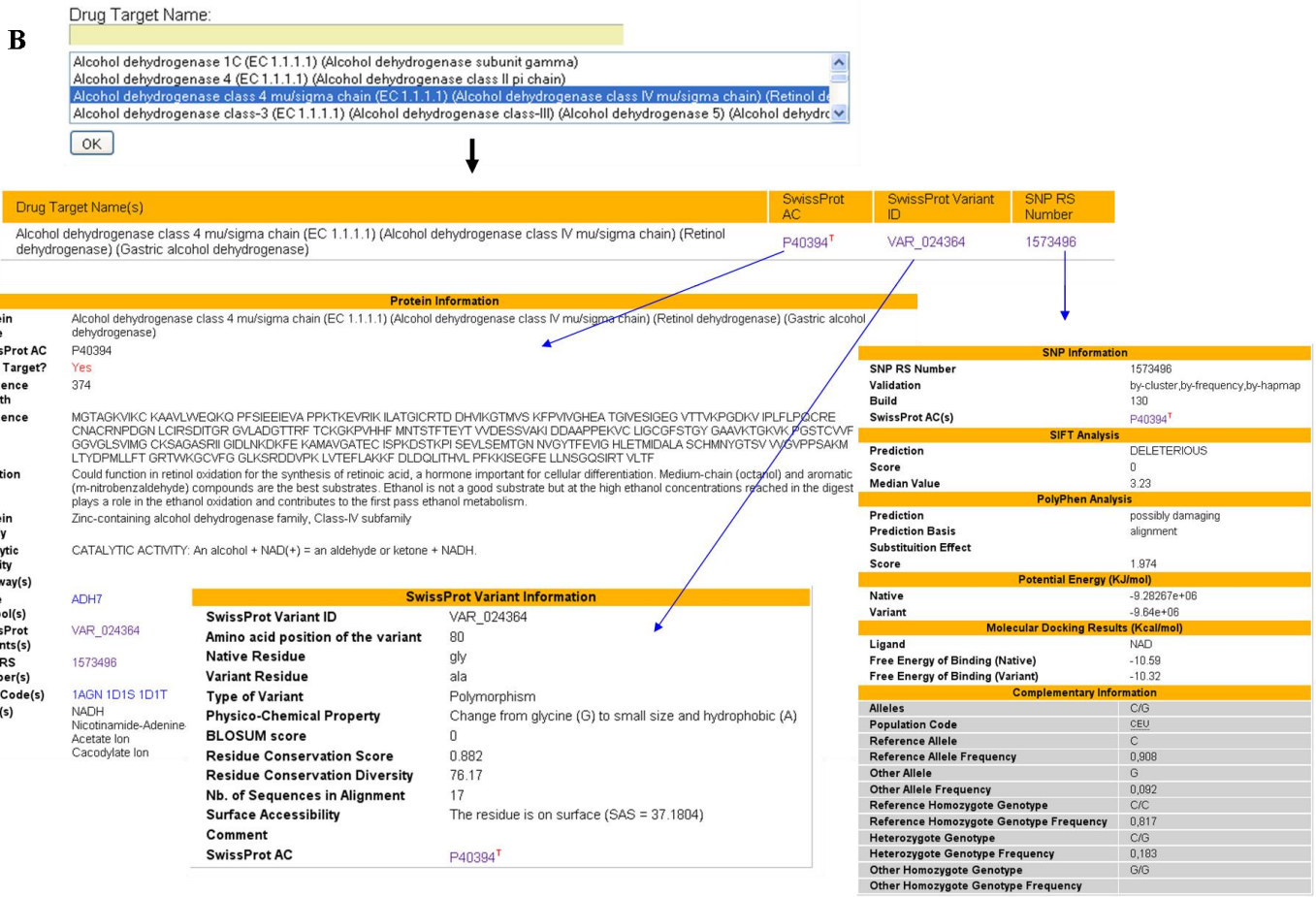
Omim info:

Omim info:

type above and/or select here
2-methyl-3-hydroxybutyryl-CoA dehydrogenase deficiency (MHBD deficiency)
3-alpha-hydroxyacyl-CoA dehydrogenase deficiency (HADH deficiency)
3-hydroxy-3-methylglutaryl-CoA lyase deficiency (HMG-CoA lyase deficiency)

OK

https://doi.org/10.26434/chemrxiv-2024-8nhvq **ORCID:** https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0



https://doi.org/10.26434/chemrxiv-2024-8nhvq **ORCID:** https://orcid.org/0000-0002-5872-7611 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0