KU LEUVEN

Integrated Bioinformatics Project 2017

SNP Effect 4.X

By: Colton Gowan, Qian Yu, Xu Xiao

Supervised by: Prof. Joost Schymkowitz, Mark Fiers, Rob van der Kant

December 16, 2017

1. Introduction

The present SNPeffect platform seeks to aggregate information from several software packages which predict the effects of protein folding or stability on experimentally derived amino-acid substitutions[1]. The next iteration of SNPeffect will accommodate large datasets. By creating a pipeline to get annotations of these substitutions, we aim to identify significant protein abnormalities in order to calculate the total 'mutation load' a cell caries in term of protein damage and better understand its role in cancer datasets.

Typically, cancer cells have many of these mutations and they lead to an upregulation of the specialized stress response pathways, mainly upregulating molecular chaperones[2,3]. Inhibition of chaperones or protein degradation is used in the hospital for treating cancer. But not all tumors respond, partly because the sensitivity depends on how many damaged proteins the cell needs to handle. Additionally, with a time series of cancer lines, may potentially identify how the mutational load changes over time and which proteins (and the locations within them) are more likely to gain mutations at earlier versus later cancer stages.

Given a variant calling format (VCF) file from an RNA-Seq data experimental group, (e.g. cancer, treatment, etc), we wish to decipher the effects of SNPs that are present within proteins. This procedure will be done through mapping, aggregating scores generated by different software, and finally querying a novel and growing database.

Mapping is performed to link the SNPs (in the VCF file) to their relevant information: genomic location, transcript IDs, novel amino acid substitution/sequence, and ultimately the PDB structure files containing this SNP. This is performed through the utilization of several databases including UniprotKB[4], RCSB, SnpEff, SIFTS, PFAM etc.

Software outputs, in the initial version, contains the following: SIFT[5], PolyPhen[6], TANGO[7], IUPRED[8], PFAM[9], and FoldX[10]. These may require the wild type amino acid sequence or the PDB files with the associated versions with the mutant amino acid. These scores are calculated through various algorithms and give us the power to discern between innocuous and destructive amino acid substitutions for normal protein behavior.

Querying from the final output of this platform will enable researchers to ask complex questions about the information aggregated into the database created from their VCF file. These queries may include specific parameters for scores from tools included in our suite. Additionally, the questions may be as specific as to the outputs of mutations within protein family domains.

Ultimately, the goal is to create an ever-expanding database which will be continuously updated as users upload new data so the number of potential queries can become more exact and have more support by public data. This protocol will be performed twice: firstly with public databases and secondly with the Switch Lab - BLAST algorithm which utilizes a synthetic PDB database, which contains PDB files for every protein, mostly through homology models.

Finally, we will discuss future paths that can use this project as foundation.

2. Methods

2.1 Workflow

All of the following was performed with the Human Genome 19 since our VCF cancer dataset was created with this version. The scripts could be updated to accommodate other versions.

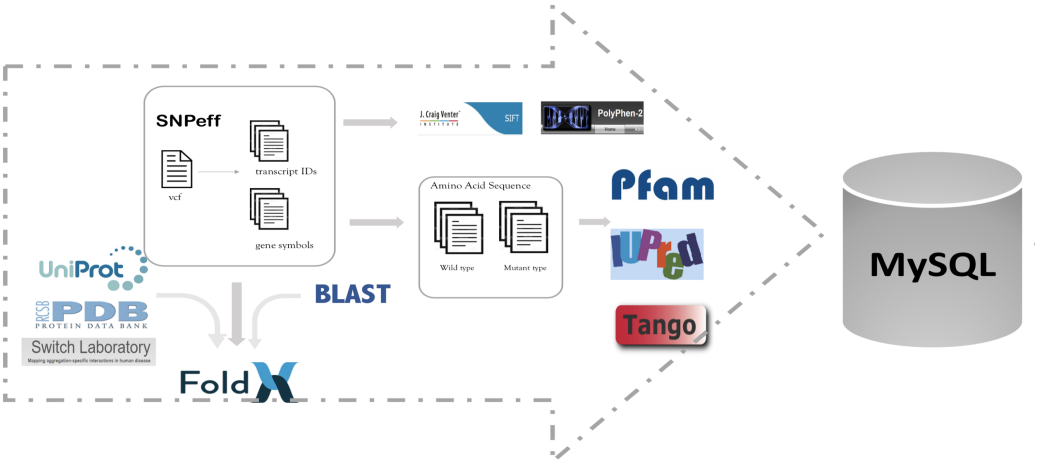

Figure 1: Workflow

2.1.1 Parsing VCF file

To discover proteins affected by each SNP, we parsed the VCF after it had been processed by SnpEff[11]. SnpEff appended an additional 'annotation' column to the standard VCF which included key information: The ENSG (Ensembl gene ID) and the associated ENST (Ensembl transcript ID) and the location of the nucleotide change within the ENST. With this information, we were able to cross-reference the ENST to the human genome annotation file to locate the exact genomic position of the transcripts.

In addition, we also used the 'Sequence Ontology term' annotated by SnpEff. With the criteria: whether this variant can alter the amino acid sequence, we filtered variants which are annotated as 'synonymous variant', 'intergenic region', 'upstream gene variant', 'non coding transcript variant', etc. (see Appendix for full name list)

The human genome annotation file was parsed to filter lines with CDS information which contained the beginning and end of each CDS region as well as the ENST IDs which contain them. Each ENST was parsed with the human genome sequence providing the wild type and mutated amino acid sequence of each transcript. Those sequence are our inputs for PFAM, IUPred and Tango.

2.1.2 Mapping genomic location to protein

Next, we integrated information from UniProt and RCSB. We used gene symbols to find corresponding UniProt IDs. To begin, we sought proteins that already had PDB files. Conveniently, a webpage on UniProt contained a complete list of UniprotKB IDs with the associated PDBs[4]. This webpage was parsed for human proteins. The output was a list of many-to-many associations between a UniProtKB IDs and PDB IDs.

Initially, we attempted to take advantage of this curated list and the SIFTS[12] database containing PDBe to UniProt residue level mapping. However, an issue arose over locating correct PDB files from RCSB database, due to many-to-many relationships between a protein and PDB file. There is no information from SIFTS that indicates which isoform is mapped to the PDB, so we couldn't map PDB to transcript level. We were unable to determine which residue in the PDB that is altered by the SNP mutation. Similarly an issue, PDB files may contain only sub-regions of the protein, missing the our SNP of interest. In lieu of this we utilized a SWITCH Lab-database which included 30,000 homology-based derived PDB files, containing homology models of about 40% of human proteins, and mapping information from the transcript level to PDB models. This allowed us to collect relevant information: the correct PDB model and residue changes due to each SNP for FoldX input.

2.2.3 Using VCF file as inputs for SIFT and PolyPhen

Without further manipulation, we also collected results from SIFT and PolyPhen, which need VCF file as inputs.

2.2 Software (See appendix)

2.3 Database

This project required an appropriate database to store each SNP's corresponding wild type and mutant annotations. There are two different database types; relational (i.e. SQL) and non-relational (i.e. NoSQL) databases. One challenge for choosing a database is that "one size does not fit all" since each type of database has strengths and weaknesses.

Traditional SQL databases are reliable and fitted for many complex application models. They support ACID(A: atomicity, C: consistency, I: isolation, D: durability) properties that make development easier. Furthermore, they have limitless indexing for fast query. However, their biggest problem is the difficulty to scale out horizontally unless they use sharding techniques, which has a limitation of 4096 columns per table. In our case the limitation is enough for the requirement; we may want to add new information from other tools for mutations in the future. NoSQL databases can be attractive due to speed, flexible data structure, and cheap built-in auto-sharding techniques. Flexible data structure makes NoSQL fast development and prototyping, but requires developers to write sensible data in it. Otherwise, it is easy to make a bad database that is challenging to manage and understand. MySQL is chosen for storing all the information and users' queries, principally for the easy use and maintenance for most developers, combined with the specific demands of this project.

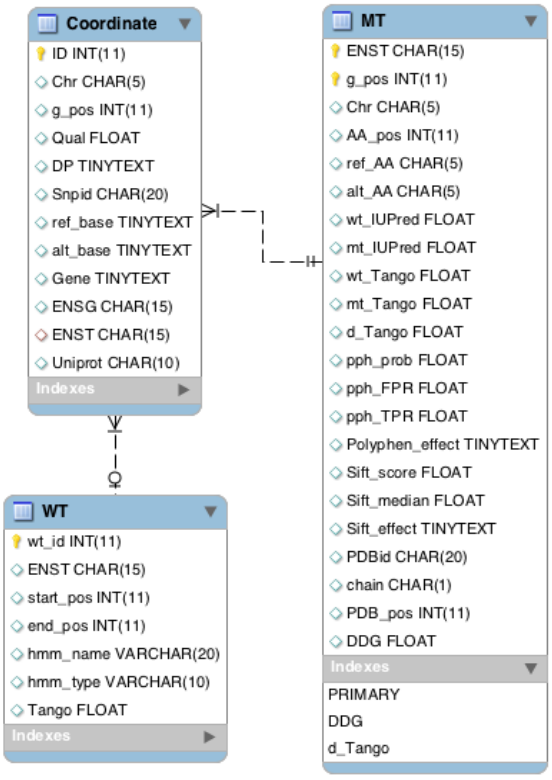

Figure 2: Database Architecture

Our database is composed of three tables: a coordinate, MT and WT table, as shown in Fig 2: database architecture. The coordinate table consists of basic SNP information from the VCF and its corresponding gene information including Ensembl genes and transcript IDs from SnpEff. The MT table stores the outputs from previous tools. For domain-specific information, the wild type table includes PFAM predicted-domains and associated TANGO scores of each domain.

3. Results and Discussion

3. 1 Results

3.1.1 Data manipulation

In our VCF file, there are 71798 SNPs in total. After annotating and filtering by SnpEff, there are 4417 SNPs left which may have impact on 2872 genes and 9457 transcripts. There are 369 variants (8.4%) which are able to be mapped to our PDB dataset. Further analysis from FoldX and Tango, 341 mutations may make protein unstable

(ΔΔG >0) and 938 mutations tend to form protein aggregation (ΔTango < -50 or ΔTango > 50). According to SIFT and PolyPhen prediction, there are 526 and 548 mutations that may damage protein respectively.

3.1.2 Database query

As FoldX calculates the effect of the mutation on the stability of the native state, expressed as a ΔΔG value in kcal/mol and TANGO predicts the aggregation propensity of the unfolded state of the protein, which typically gives a zero score for most residues, expect for short aggregation prone region (APR) with a score of up to 100, we may want to find all mutations that make ΔΔG > X in domain with TANGO > Y. Since most residues are thus not part of an aggregation prone region, mutations that affect the APRs directly are rare, but a large fraction of disease mutations are destabilizing and hence cause the exposure of APRs, promoting aggregation in that way[13,14].

Snpid	ref_AA	alt_AA	ENST	Gene	Uniprot	Tango	DDG	hmm_name
rs1071816	Y	F	ENST00000412585	HLA-B	P01889,	263.905	5.33411	MHC_I
NULL	V	A	ENST00000377698	ALDH1B1	P30837	2042.36	5.09428	Aldedh
rs11890	G	V	ENST00000219302	NME3	Q13232	903.451	5.68082	NDK

Figure 3: Query result – all mutations that make ΔΔG > 5 in domain with TANGO > 0

3. 2 Discussion

Our work has resulted in a single database that pulls information for each SNP from various tools. These tools require different inputs which our pipeline creates from an initial VCF file. The culmination of data may allow researchers interested in mutant proteins to quickly filter out mutations of interest based on queries made in a novel database.

Since each tool may have multiple outputs which factor into a possible query, they may become very complex. The type of query may depend on the background of the researcher, what specific questions they are interested in, and how well they understand each of the outputs. On the other hand, it may be possible to use a machine learning approach to create successful queries. This may potentially utilize many outputs at once.

Time series data from cancer lines may be analyzed by comparing the resultant databases from each source. This has the potential to reveal patterns in locations and types of proteins which become mutated with different rates.

Mutation load, in this context, refers to the proteome level of mutations, factoring in expression. Researchers are interested in how proteins containing cancer-derived mutations are over-expressed what the physiological events occurring on the cellular scale. This may include protein-causing aggregates becoming overexpressed, potentially limiting the activity of pharmaceutical action to prevent cancer growth.

Finally, the database design should allow for information from other sources to be continually added as they become available. This means that with our example dataset, the number of elements associated with each SNP could easily be expanded. Another option is to pre-calculate the SNPs in protein sequences and upload them to a web-based database so researchers can quickly query the results without requiring computation time for the various algorithms.

Appendix

Scripts address: **https://github.com/yuq1993/IBP-SNPeffect**

Screen list:

Here is the list of 'Sequence Ontology term' annotated by SnpEff. We filtered out those SNPs.

Non-coding transcript variant
Splice region variant & intron variant
5 prime UTR variant
Intergenic region
Synonymous variant
Non-coding transcript exon variant
3 prime UTR variant
Upstream gene variant
Downstream gene variant
Intron variant
Splice donor variant & intron variant
Splice acceptor variant & intron variant
Sequence feature
5 prime UTR premature start codon gain variant
Splice region variant
Intergenic region
Structural interaction variant
INDEL
Splice region variant & non-coding transcript exon variant
Protein protein contact
Splice donor variant & splice region variant & intron variant
Intragenic variant
Splice region variant & synonymous variant

Software:

In this section, we will provide a more detailed description of the suite of tools included in our platform. As well as the outputs included in our final database.

"TANGO is based on the physico-chemical principles of beta-sheet formation, extended by the assumption that the core regions of an aggregate are fully buried"[7]. The overall goal of TANGO is to detect regions that promote protein aggregation. Our outputs included in our database from this tool were the TANGO score and deltaTANGO. The TANGO score is calculated for each residue which may be summed over the entire sequence or a subregion. We are able to calculate the TANGO score and a deltaTANGO to compare the wild-type and mutant sequences.

"Polyphen-2 PolyPhen-2 uses eight sequence-based and three structure-based predictive features which were selected automatically by an iterative greedy algorithm … The functional significance of an allele replacement is predicted from its individual features by Naïve Bayes classifier"[6]. This software includes 3

outputs, all of which are included in the final database: TPR, FPR, and an overall classification being benign or possibly/probably damaging.

“SIFT (Sorting Intolerant From Tolerant) is a program that predicts whether an amino acid substitution affects protein function so that users can prioritize substitutions for further study”[5]. The algorithm is based on sequence homology to see if mutations occur in highly conserved residues ; no structural data is directly used to infer a SIFT score. The three outputs are included in our final database : a SIFT median (informs about the diversity of sequences used in the calculation, high value indicates score generated based on closely related sequences), a SIFT score (predicted amino acid substitution damage, quantitative) and a SIFT prediction (overall decision of the degree of damage , qualitative)

IUPRED seeks to predict regions which are classified as the following: “ proteins and domains (IUPs) [that] lack a well-defined three-dimensional structure under native condition ”[8]. Such mutations causing an increase in the IUPRED score may be more likely in causing issues related to cancer cells. Similar to TANGO, an IUPRED score is calculated for each amino acid residue which may be summed up over an entire sequence or subregion.

“FoldX is an empirical force field that was developed for the rapid evaluation of the effect of mutations on the stability, folding and dynamics of proteins and nucleic acids. The core functionality of FoldX, namely the calculation of the free energy of a macromolecule based on its high-resolution 3D structure...”[10]. FoldX utilizes the PDB files. It contains the most detailed information of biochemistry to predict the functional relationships between individual. There are many outputs included in the raw output. In this initial version of our database, we’ve included the delta-Intrinsically unstructured/disordered proteins and domains (IUPs) lack a well-defined three-dimensional structure under native conditiondeltaG value which summarizes the results in a single value, although more could easily be included.

”Pfam is a comprehensive collection of protein domains and families, with a range of well-established uses including genome annotation”[9]. We utilized the PFAM database to locate SNPS contained within protein domains. This was an important aspect of our project because it allowed us to use the previously listed software, to specifically check potential changes within these domains. This is of significant importance because these domains are highly structured in the tertiary protein structure and changes within them have significant impacts, compared to the regions between such domains.

References

1. De Baets, G. *et al.* SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants. *Nucleic Acids Research* **40,** D935–D939 (2011).
2. Dai, C. & Sampson, S. B. HSF1: Guardian of Proteostasis in Cancer. *Trends Cell Biol.* **26,** 17–28 (2016).
3. Tang, Z. *et al.* MEK guards proteome stability and inhibits tumor-suppressive amyloidogenesis via HSF1. *Cell* **160,** 729–744 (2015).
4. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Research* **45,** D158–D169 (2017).
5. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research* **31,** 3812–3814 (2003).
6. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat Meth* **7,** 248–249 (2010).
7. Fernandez-Escamilla, A.-M., Rousseau, F., Schymkowitz, J. & Serrano, L. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.* **22,** 1302–1306 (2004).
8. Dosztányi, Z., Csizmok, V., Tompa, P. & Simon, I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21,** 3433–3434 (2005).
9. Bateman, A. The Pfam protein families database. *Nucleic Acids Research* **32,** 138D–141 (2004).
10. Schymkowitz, J. *et al.* The FoldX web server: an online force field. *Nucleic Acids Research* **33,** W382–8 (2005).
11. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* **6,** 80–92 (2014).
12. Velankar, S. *et al.* SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Research* **41,** D483–D489 (2012).
13. Ganesan, A. *et al.* Structural hot spots for the solubility of globular proteins. *Nature Communications* **7,** 10816 (2016).
14. Rousseau, F., Serrano, L. & Schymkowitz, J. W. H. How evolutionary pressure against protein aggregation shaped chaperone specificity. *J. Mol. Biol.* **355,** 1037–1047 (2006).