Protein structure modeling for variant pathogenicity prediction

Author: Sylt Schuurmans
Studentnumber: 333332
Study: Bio-informatica
Hanze hogeschool: Institute for Life Science and Technology
Hospital: Universitary Medical Center Groningen: Department of Genetics
Supervisor : Joeri van der Velde

June 7, 2019

Introduction

Around 1 in 17 people is affected by one of 7,000 known rare diseases. Most of these patients do not receive a diagnosis, which means they remain in uncertainty without a prognosis, are unable join specific patient support groups, and do not receive the most appropriate treatment. Next-generation sequencing (NGS) of DNA promises to establish a molecular diagnosis and help these patients but many challenges still stand in the way of maximum success. Recent years have seen great advances in computational tools that quickly reduce the amount of DNA variants to be interpreted by a human expert for potentially pathogenic effects []. Although algorithms can now safely remove around 95% of the harmless variants, this still leaves hundreds of variants to be investigated for a whole-exome sequenced patient, which is far too many for a quick and clear diagnosis. Current tools to predict variant pathogenicity rely on features such as evolutionary conservation, annotation of regulatory genomics elements or structural DNA features. These tools have already been optimized over many years and further significant improvements are not expected. Therefore there is still a great need for even more powerful variant prioritization tools. A refreshing alternative was presented by VIPUR [] which shows the potential of structural modeling of proteins to predict the actual effect of a specific variant on the function of that protein. This presents an exciting new opportunity to improve genome diagnostic variant prioritization. However, this predictor was (i) not integrated with the latest and greatest variant pathogenicity prediction approaches, (ii) was trained on relatively small number of variants, and (iii) did not result high quality software that was ready to be taken into routine diagnostic practice. To test this approach we will explore the potential pitfalls of protein modeling by evaluating the VIPUR pipeline and by examining a single protein with it variants.

Abbreviations

3D Three Dimensional
ACCP Solvent Accessible Surface Area
API Application Programming Interface
Bash Bourne Again Shell
CPU Central Processing Unit
CSV Comma Seperated Values
DNA Deoxyribonucleic Acid
DISC Death-Inducing Signaling Complex
FADD Fas Associated Death Domain protein
FasL Fas Ligand
FEM Fixed End Move
FHF Familial Hibernian Fever
GAVIN Gene-Aware Variant INterpretation
GRCh/hg Genome Reference Consortium Human Human Genome
HOPE Have yOur Protein Explained
LOMETS Local Meta-threading Server
MD Molecular Dynamics
MPI Message Parsing Interface
NCBI National Center for Biotechnology Information
NF-κB Nuclear Factor kappa-light-chain-enhancer of activated B cells
OpenGL Open Graphics Library
OS Operating System
OSF Open Science Framework
PDB Protein Data Bank
PM Pivot Movement
PSI-BLAST Position Specific Iterative BLAST
PSSM Position Specific Scoring Matrix
RCSB Research Collaboratory for Structural Bioinformatics
REU Rosetta Energy Unit
RNA Ribonucleic acid
RSMD Root Mean Square Deviation
SASA Solvent Accessible Surface Area
SCOP The Structural Classification of Proteins
SLURM Simple Linux Utility for Resource Management
SODD Silencer of Death Domain
SPVAA Simple Protein Variant Analysis Approach
TNF Tumor Necrosis Factor
TNFR1 Tumor Necrosis Factor Receptor Superfamily Member 1A TNFRSF1A Tumor Necrosis Factor Receptor Superfamily Member 1A
TRADD Tumor Necrosis Factor Receptor type 1-Associated DEATH Domain protein
TRAPS Tumor Necrosis Factor Associated Receptor-Associated Periodic Syndrome
VIPUR Variant Interpretation Using Rosetta
VTS VIPUR Training Set

Contents

List of Figures

List of Tables

1 Variant prediction in genome diagnostics and the addition of protein modeling

1.1 Mutations and its effects in the central dogma of molecular biology

Within the human genome mutations occur continuously by internal and external factors that substitute, remove, insert or alter the reading frame in a nucleotide sequence. Mutations are not without consequences and can be: beneficial, benign or in most cases pathogenic because they replace a nucleotide which serves a purpose at the specific position in a sequence. Alterations in sequences might lead to a difference in ribonucleic acid (RNA) transcription rates or differences in the RNA transcript that is formed from the deoxyribonucleic acid (DNA) which both can influence the cellular machinery. Mutations outside a gene could lead to lowered or heightened transcription of a protein, when a mutation resides inside a gene it could lead to proteins that are: unstable during or after formation, perform less optimal or are not functional [].

1.2 A general concept of structural levels within proteins and the effect of mutations

The formation of protein structures is classified in different levels, distinctions are made based on bindings and structures that arise from them. The order in which amino acids appear in a sequence is called the primary structure, in this level amino acids are only bound to each other by peptide bonds. Within a primary structure amino acids can form new peptide bonds between the N and C -terminus of an amino acid, with these bonds 3D structures are made called α helices and β-sheets that together make up the secondary structure. More alterations to a single amino acid sequence in the 3D can come from disulfide bridges, ion, hydrogen -bonds, hydrophobic and hydrophilic -interactions formed by the residues of the amino acids, together these bonds form the tertiary structure. By combining multiple tertiary structures the quaternary structure of a protein can be formed out of the mentioned bonds, bridges and interactions [].

Mutations within proteins can have different effects to protein structures, often single missense mutations often have minimal effect on the backbone of a protein [] but can result in destabilization of the structure when assembled or can disrupt the active site. Frameshift mutations often cause larger disruptions within the structure and often lead to proteins that are deformed or have early stop codons [].

1.3 Addition of structural data to diagnosis and treatment in healthcare

Acquiring information about DNA sequences highly relies on experimental sequencing methods and became cheaper over the years [] and found its use in diagnosing patients within the healthcare sector []. From the collected data by genome sequencing experiments most of the analysis is handled in-silico due to the quantities of data that is produced. Proteins often find their use in diagnosing diseases experimentally [], however in-silico it is often limited to information about conservation in the amino acid sequence which may lead to identical results as by analyzing DNA. Yet the 3D structure defines how a protein functions [] and by assessing structures of protein variants it becomes possible to determine the change in function and diagnose protein variants that were unclassifiable through finding conservation. Another advantage of the structural information is that it gives the possibility to develop treatment for diseases that are caused by a mutations. With experimental methods such as X-ray crystallography and nuclear magnetic resonance (NMR) more than 158000 structures [] have been completely revealed, however it is only a tiny fraction of the potential proteins possible without folds []. Making 3D structures is currently not common for diagnosis because it is relative expensive and is difficult to perform, some structures contain flexible regions which makes it hard to determine their exact position and can cause information loss about the structure [].

1.4 Protein modeling techniques

An alternative approach to determine structures is based on modeling the protein structure computationally from the amino acid sequence of the desired protein. A downside from computer generated models is that they do not follow the rules of physics and therefore not automatically fold into the correct confirmation. With the method homology modeling sequences of the requested protein are aligned to sequences of known experimental determined structures, based on these alignments a template is formed whereon structural fragments are built, it is not recommended to use this strategy if the sequence identity is less than 20% since there might not be any structural relation at that point []. Another approach is protein threading which relies on the observation of folds in previous determined experimental structures. Based on the occurrence of specific folds a probability is predicted that a certain residue in a protein might fold in that manner.

Strategies are continuously being improved and developed for proteins to determine the unknown structures, but all have the similar guidelines in avoiding steric hindrance [] and finding the lowest energies based on different scoring systems []. From the computer generated models many are less accurate than the experimental determined methods and are often compared to them for reference. However the computational models do not have follow the same laws of physics which bottleneck the current experimental methods in for example determining membrane proteins [].

1.5 A theoretical large scale implementation of structural protein variant assessment

With the wide spectrum of potential different proteins it can be difficult and maybe momentarily impossible to produce any form of universal protein assessment standardization that is able to determine if a mutation is harmful or not based on structural information. However a first step to solve such a complex problem would be by determining the correct approach, in this case it is assumed that a machine learning approach would be the best method for detecting patterns in structures and classifying the effect of structural changes. Because it has the ability to learn from structural mutations currently available, assuming that the current knowledge about structures and mutations is correct, and is able to develop new insights in how structural changes could affect proteins.

Since the problem is so complex it should be divided into smaller more feasible problems, beginning by separating the different protein classes, which for example can be done according to The Structural Classification of Proteins database (SCOP) []. A first discrimination between the proteins could be made based on protein type/fold class (membrane, globular, fibrous and disordered -proteins) because these differences already predetermine some functions and locations for certain proteins in a cell []. After formation of these classes each should have its own machine learning method applied so their features can be analyzed within context of where and how they function. The next set of discriminators is highly dependent on the variations in classes, but all have features in the end describing bonds, interactions and movement of complexes in protein structures. When for each of the main classes a method has been developed a meta classifier will determine based on certain aspects which method should be applied to determine the effect of mutation in a protein.

2 Monte Carlo method

2.1 Monte Carlo method

There are complex problems in a variety of research fields which could take up years or even centuries to compute with simple deterministic methods. For some problems there is an algorithm which makes it possible to cut down computation time significantly, but when no deterministic algorithm is available to speed up the process an empirical probabilistic method might be able to approximate the desired result. With the Monte Carlo method random samples are taken from the parameter space ,that describe a data set, and fed into a model which produces a potential outcome. By repeating the process more results are generated until at some point the data can display a pattern that describes the outcome. The result is a quantified probability which describes the chance that something might occur based on the quantity of occurrence generated by the model [].
The Monte Carlo methods can differ depending on the algorithm and application in which it is used, but in summary most implementations will follow a general pattern []:

0. Construct a model which is able to describe an outcome of the problem.
1. Define the space of which inputs can be used by the model to get an outcome (creating a parameter space).
2. Use the model to generate results based on random sampled input from the parameter space.
3. Order and determine which results are part of a certain outcome and draw conclusions on the generated statistical evidence.

2.2 The use of the Monte Carlo method and its pitfalls

The Monte Carlo method is widely used within various applications in different fields of science but it is limited in the type of problems it can solve and is suitable for; problems of which all the inputs are known but it is too inefficient to compute deterministically; situations that require uncertainty to be incorporated into the analysis and exploring parameters for a model that give a better impact than the current parameters. The mentioned type of problems it can solve all tend to rely on significant quantities of data which makes it a relative time consuming process for generating results. Meaning of the generated result is highly depended on the model and random sampling techniques which both contribute to an errors in the result [].

3 TRAPS disease and its proteins

3.1 Tumor Necrosis Factor Receptor Associated Syndrome

Tumor necrosis factor receptor-associated periodic syndrome (TRAPS) is classified as a rare disease (1 : 1,000,000) and was formerly known as Familial Hibernian fever (FHF) [], is a hereditary autosomal dominant disease which can cause recurring fevers with a duration from days up to several months. Symptoms during these fevers are: skin rash, swelling, inflammatory reactions across the whole body and pain in the abdomen, muscles and/or joints, a long term and lasting effect is the accumulation of amyloid within the kidneys and may result in other diseases []. TRAPS is known to be caused by mutations within the gene turmor necrosis factor receptor 1 (TNFRSFA1/TNRF1) (Section 3.2), the mutated proteins tend to get trapped in the cell and will be unable to reach the cell surface and therefore start activating a inflammatory response []. So far 158 mutations have been associated with the disease [], but more mutations have been identified in TNFRSF1A wherein some might be pathogenic (Sections 4.10, 4.11).

3.2 Tumor Necrosis Factor Receptor Super Family Member 1A

Tumor Necrosis Factor Receptor Super Family Member 1A (TNFRSF1A, TNFR1) is a gene located on chromosome 12 region 1 band 3 and sub-band 31. The gene produces a trans-membrane receptor consisting of 445 residues divided into 221 residue cytoplasmic section and a 171 extracellular part that consists of 4 conserved cysteine rich domains []. The receptor is ubiquitous across most cell surfaces ,but not on erythrocytes [], and can form two different types of unbound hexagonal clusters depending on the dimer formation []. When the structures are dimers the binding sites are exposed and make it possible for tumor necrosis factor (TNF) α and β (Section 3.3) to bind in trimeric form, with binding of TNF the dimers disconnect and three TNFR1s interact with the TNF trimer []. With the interaction of the TNF trimers with TNFR1 it can activate several pathways such as; the nuclear factor kappa-light-chain-enhancer of activated B cells (NF-κB), which enhances the transcription of various genes during inflammation, infection or other forms of external stress; also it is able to activate the extrinsic pathway of apoptosis after binding of TNF to TNFR1, by releasing the silencer of death domain (SODD) proteins release on the cytoplasmic site. Tumor Necrosis Factor Receptor type 1-Associated DEATH Domain protein (TRADD) [] will start to bind together with proteins that will form a complex which will attract Fas associated death domain (FADD) and after two hours[] if not inhibited. On binding of FADD initiator caspase 8 starts a cascade wherein caspase 3 is activated an will cleave aspartate out of proteins and thereby disrupting the metabolism[].

3.3 Tumor Necrosis Factor Alpha and Beta

The proteins TNF α and β are both pro-inflammatory cytokines that are produced as response to an infection or when a cell is damaged. Both are transcribed from their genes that reside in chromosome 6 in the p-arm at region 2 band 1 and sub-band 3. TNF α and β are 35% identical and 50% homologous to each other consisting out of 233 and 205 amino acid residues. Both are able to form a homotrimeric structures that can bind to the dimeric form TNFR1 (Section 3.2) to activate the extrinsic pathway[].

4 Materials and methods

4.1 VIPUR approach

Variant interpretation using Rosetta (VIPUR) is a machine learning approach for predicting deleteriousness of proteins (loss of function) and uses sequential and structural information. To train VIPUR a training set was made that contains sequence and structure features. Structures within the VIPUR training set (VTS) were collected from Modbase [] and SWISS-MODEL []. Proteins that did not have a structure within the VTS were modeled with Modeller (Section 4.9 based on protein fragments that had the highest amino acid sequence identity to the protein. Some structures from the databases had: duplicate chains, ligands, metals and non-standard amino acids which were removed to avoid inconsistencies that could alter the features generated by the tools and hinder feature collection within Rosetta tools (Section 4.3), therefore they were removed to make the data homogeneous. For all proteins a mutation file was made that described where and which residues had to be mutated, with this file DDG monomer (Section 4.3.2) can predict changes in the protein structure of which all features are used for prediction. Structural mutations of proteins that are in the VTS were introduced by a script using PyMOL (Section 4.15) by default or PyRosetta (Section 4.4) if PyMOL was not available. After mutation the structure is optimized by the relax application (Section 4.3.1) and produces 50 relaxed structures of a single variant where the properties of each protein are written to a score file of which the quartiles are used as a learning feature. Probe (Section 4.6)calculated the solvent accessible surface area (SASA/ACCP) of a protein in square Ångstrom ($Å^2$) which is a structural machine learning feature. The sequence features of VIPUR are produced by PSI-blast (Section 4.5) on non mutated sequences and blasted against the NCBI protein database (nr) which results in a position specific scoring matrix (PSSM). From the PSSM scores of the non-mutated, mutated, the difference in scores between non-mutated and mutated, information content and the difference between groups [] are all sequential features for VIPUR. With 106 features generated by the mentioned tools deleteriousness of a protein variant is determined with sparse logistic regression. The term sparse implies that a limited set of features was used because the weights "shrink" to 0 with regularization [].

4.2 Simple protein variant analysis approach

Another method for determining function loss in a protein variant is through assessment of difference in energy levels between a wild type and a variant in the complex where it resides. Analyzing mutations from this perspective gives the ability to view a complex in whole and determine how residues cause perturbations in a complex. To make a variant of the wild type, a structure was required wherein a missense mutation could be introduced with Modeller (Section 4.9). The backbone structure of the variant was modified with the backrub application (Section 4.3.4) to make it better interact with other amino acid backbones in the structure resulting in 1000 models. The lowest Rosetta scoring (Section 4.3) one would be selected to further improve the side chains with the Relax application (Section 4.3.1) which makes the side chains of the protein move into lower energy formations. From these 64 models were made where its energies levels were compared to the native structure to determine the effects a mutation would have on the protein, the lowest scoring ones were used as figures. This method shows similarities to that of VIPUR and was only tested on TNFRSF1A (Section 3.2) and its ligands TNF α and β. This method keeps: duplicate chains and protein ligands within the structure, water is excluded since it can cause issues with Rosetta tools (Section 4.3). This procedure with these steps is defined as the Simple protein variant analysis approach (SPVAA).

4.3 Rosetta

Rosetta is a software suite that has a variety of tools that are developed to aid in macro molecular and antibody ,analysis, design and prediction []. However no tools in the suite have been encountered that could introduce missense mutations in the proteins and has been dealt with by other software (Sections 4.15, 4.4, 4.9). With the introduced mutations water had to be removed because some tools cannot

predict structures well with: water, metals and amino acids that are no part of the standard (20) amino acids [].

Within the tools from Rosetta various scores are assigned to different properties related to bonds, interactions, energies and geometries within structures and are written to a score file. The scores can be used to compare multiple protein variants with each other which can be done based on the the "total_score", Rosetta score or Rosetta energy unit (REU). REU is based on a combination of scores and favors low energy models, but it also has statistical terms influence the score based on known favorable folds from existing structures that reside in the curated Rosetta database [].
Rosetta software suite Version 3.10

4.3.1 Relax

The Relax application was used by VIPUR and by SPVAA to relax the side chains to minimize energy levels within the local conformational search space [] of the structure, it determines the most likely energy levels with the Monte Carlo method (Section 2.2) and after a certain set of moves it produces a structure and stars anew[]. Scores from each produced by Relax were written into a single score file.

4.3.2 DDG Monomer

DDG monomer is meant to predict energetic stability of a point mutation in monomeric protein.The application was used by VIPUR to collect features related to energies and hydrogen, disfulfide, bonds and constraints differences between the wild type and a mutated protein. To execute the tool a script had to be ran that renumbers the wild type pdb file and it requires a "mutation file" that describes the change of a residue based on name and position changes to a different residue [].

4.3.3 Rescore

With this tool Rosetta scores can be calculated based on silent or PDB files proteins structures [] , the output is identical to that is written within the score files produced by Relax (Section 4.3.1).

4.3.4 Backrub

The backrub application is based on the Monte Carlo method (Section 2.2), and alters a protein by moving its backbone residues with a strategy called fix end move (FEM). With this strategy, groups of residues are selected at random from the structure, it can contain up to: four dihedral, two bond angles and two end points. Both ends of a group are fixated at their position in which a new angle α arises, within this angle residues are pivoted in their natural occurring maximum range of $\pm 10°$ []. With this method the backbones of newly introduced mutations were altered, for each attempt a new file was generated and a score was written to a score file, from which the lowest Rosetta scoring was selected to be further relaxed (Section 4.3.1). It was used on the mutated protein to relax the modified backbone structure.

4.4 PyRosetta

Is an application programming (API) which has Python bindings (Section 4.18) for the Rosetta software suite (Section 4.3), it founds its use in VIPUR when no PyMOL (Section 4.15) was available to mutate residues within a structure [].
Version 4

4.5 PSI-BLAST

Position specific iterative basic local alignment search tool (PSI-BLAST) focuses on distant relatives of proteins by making a profile of the sequence and querying it at a protein sequence database. With the generated results a new profile is constructed and queried again, these steps are repeated several

times to determine which residues are found in relatives of the protein. The result is a position specific scoring matrix (PSSM) which describes the frequency of which residues are substituted by a specific other residue, positive is more, negative is less common []. From the PSSMs sequences features were acquired for the VIPUR machine learning method.
Position-Specific Iterated BLAST 2.7.1+

4.6 Probe

Probe is able to evaluate atom packing for a single protein or interacting proteins by creating a probe, which is described as a sphere like object, that marks an area with dots when at least two non-covalent atoms are in contact with the probe at the same position []. VIPUR used this tool to calculate solvent accessible surface area (SASA or ACCP).
version 2.16.130520

4.7 Robetta prediction server

The web tool Robetta integrates several tools to form protein structures based on sequence alignments of previously discovered structures also known as homology modeling (Section 1.4). It requires an amino acid sequence, optionally constrains and fragments can be added to disallow movement of certain structures or add known fragments to avoid calculating pieces that are already known. With this information Robetta search with the help of sequence aligners for known fragments and tries to incorporate them into a single protein structure [].

4.8 I-TASSER prediction server

The I-TASSER web server is a tool that is able to predict protein structures with a FASTA sequence. The first step it takes is finding structural templates which resemble the sequence by local meta-threading server (LOMETS). LOMETS starts with multiple sequence alignment of which several sequences will undergo protein threading by different programs to form structural templates. The templates are assessed based on the highest alignment Z-score, the program specific confidence score and sequence identity []. The known fragments of TNFRSF1A (Section 1.4) were given as a template to I-TASSER and modeled into a whole protein to make it possible to introduce mutations and predict pathogenicity of a variants.
Server version

4.9 Modeller

Modeller is software that is developed for homology modeling but it was used for its utilities which allowed to; complete protein data bank (PDB) structures with missing atoms; predict disulfide bonds that were missing and mutate protein residues [].
Version 9.21

4.10 GAVIN Machine Learning Data Table

Is a collection of nucleotide mutations from rare diseases used by the GAVIN [] machine learning approach. From this set the genes of TNFRSF1A (Section 3.3) with a missense mutation were filtered (Section 4.19) and written into a format which the variant effect predictor could (VEP) [] could read and translate from nucleotide to protein mutations. The classification of these variants was according to Clinvar significance values [].

4.11 GenomAD

The GenomAD database consists of unified data from large scale genome sequencing data projects and is based on genome reference consortium human genome build 37 human genome 19 (GRCh37/hg19).

From this database missense mutations were collected for TNFRSF1A (Section 3.2), no classification was known for these mutations [].

4.12 Infevers

Is a website about hereditary auto immune diseases with for each disease a downloadable table about the known mutations and their classification, when classified with pathogenic or benign its function has been observed. The table for TRAPS disease (Section 3.2) was used to collect missense mutations of TNFRSF1A gene [].

4.13 Research Collaboratory for Structural Bioinformatics

Research Collaboratory for Structural Bioinformatics (RCSB) is a database where whole or fragmented experimentally determined proteins structures that are published can be found and downloaded. The Fragments for modeling (Sections 4.7, 4.8) whole TNFRSF1A (Section 3.2) (1EXT []) and determining the differences in energy levels (Section 4.3.1) with TNF β (1TNR []) with the interaction site were acquired from this database [].

4.14 Uniprot

Knowledge from various omic domains about proteins has been linked together into single database called Uniprot which makes all information accessible at once, for TNFRSF1A (Section 3.3) the FASTA sequences were collected from Uniprot and for structures it redirected to (Section 4.13) [].

4.15 PyMOL

Visualization of 3D structures, making images of proteins, putting the known orientations of monomeres in position, replacing TNF β with TNF α and aligning the structures to measure the distance between X-ray crystal structures and the produced models of were done in PyMOL []. PyMOL had a different use in VIPUR where it was used in combination with Python (Section 4.18) to perform mutagenesis on the protein structures to introduce a missense mutations.
Version 2.2.3

4.16 HOPE

Have yOur Protein Explained (HOPE) is a web service that collects information of about a user specified missense mutation in a protein and comes from various sources. Uniprot (Section 4.14) is queried with BLAST to find homologous sequences and structures, other features that are found on Uniprot are active sites, domains and various other sequence features that help to identify the function of a region. From the BLAST results homology models are made with Yasara that are sent of to WHAT IF web services that calculate structural information about the protein. Before the formation of a report all information is put into a decision tree to asses mutational effects in contexts of: contacts, structural locations, non-structural features, previous variant information and amino acid properties to form an automated report []. With this method it is not possible to asses ligands and complexes at once but only a single missense mutation within a monomer. *Version 1.1.1*

4.17 Bash

Unix like operating systems (OS) have a shell which allows users to interact with programs on a computer or with the computer itself based on commands submitted. The default shell for MacOS and also for several Linux distributions is the Bourne again shell (Bash) which was used to launch Python scripts (Section 4.18) and submit jobs to the SLURM workload manager (Section 4.20).

Laptop Version GNU bash, version 3.2.57(1)-release (x86_64-apple-darwin18)
Server Version GNU bash, version 4.1.2(2)-release (x86_64-redhat-linux-gnu)

4.18 Python

Both VIPUR and the pipeline that minimizes backbone (Section 4.3.4) and side chain energies (Section 4.3.1) were written in Python due to its capabilities, ease of use and because modeller (Section 4.9) for MacOS relies on the system version of Python and does currently not support newer versions besides the one found within the OS of Mac. The mutations that were put together from the different tables (Sections 4.11, 4.12, 4.10) with R Section 4.19) were filtered by a Python script. To apply each mutation correctly on the proteins in the detailed method a script was written in which files were generated that described in a compact format on which chains and position a mutation resided.
Laptop version 2.7.15
Server version 2.7.11

4.19 R scripting language

With R the tables from GenomeAD, GAVIN and Infevers (Sections 4.11 4.10 4.12) of TNFRSF1A missense mutations (Section 3.2) were merged together in a new comma seperated values file with their known classifications. Ordering and filtering the double mutations and removing double classifcations where done with Python (Section 4.18). It has also been used in combination ggplot2 and data.table to make density plots of all scores acquired from Rosetta Backrub and Relax.
R scripting front-end version 3.5.2 (2018-12-20)

4.20 SLURM

For computational jobs where a laptop or desktop does not suffice because due to the lack computational resources a computer cluster could come to aid. These clusters consist out of several computers that execute resource intensive tasks, to manage these systems for many clients and to use these clusters optimal a workload manager mlike simple Linux utility resource management (SLURM), is installed. Jobs are submitted that request resources for execution and are scheduled on the systems queue which is ordered based on priorities, resource requirement and time.

4.21 MPI

Some tools from the Rosetta software suite (Sections 4.3) have the ability to use multiple central processing unit (CPU) cores from a single computer or from multiple computers. With a message parsing interface (MPI) it is possible for software to communicate between CPU cores on the same and on different computers to exchange information about processes and therefor solving solutions faster.
OpenMPI/1.8.8-GNU-4.9.3-2.25

5 Results

5.1 Reviving the VIPUR approach to expand rare disease diagnostics

5.1.1 Preparatory steps for using the VIPUR approach

After the publication of VIPUR the tools, data and applications became available at the open science framework (OSF) [] which were downloaded and reviewed. All applications from the Rosetta software suite (Section 4.3) were pre-compiled without support for MPI (Section 4.21) and with that not the ability to benefit from multiple CPUs. The Rosetta software suite was rebuilt with MPI support in a slurm job where the compilation could benefit from multiple CPU cores.

5.1.2 VIPUR resolving system incompatibilities

Within the VIPUR pipeline residues were mutated to determine the effects of a structural mutation, by default missense mutations were inserted with PyMOL (Section 4.15), an alternative method integrated within the pipeline for situations wherein PyMOL was not accessible Pyrosetta (Section 4.4) could be used. Neither of these programs could be built or compiled because the lack of Open graphics library (OpenGL) for PyMOL and having the incorrect C++ and C libraries for PyRosetta. To bypass both programs and still be able to introduce mutations into PDB files Modeller (Secton 4.9) was introduced and built.

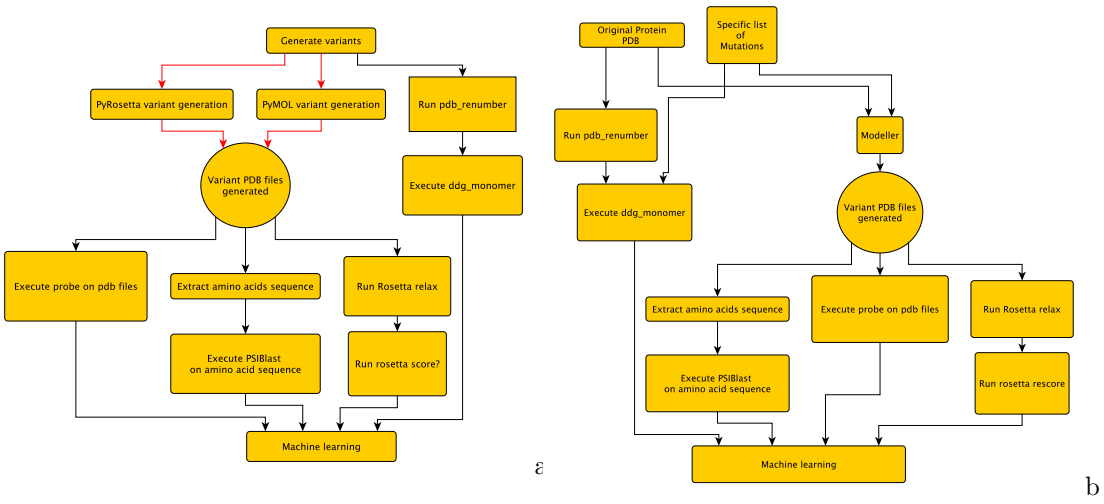
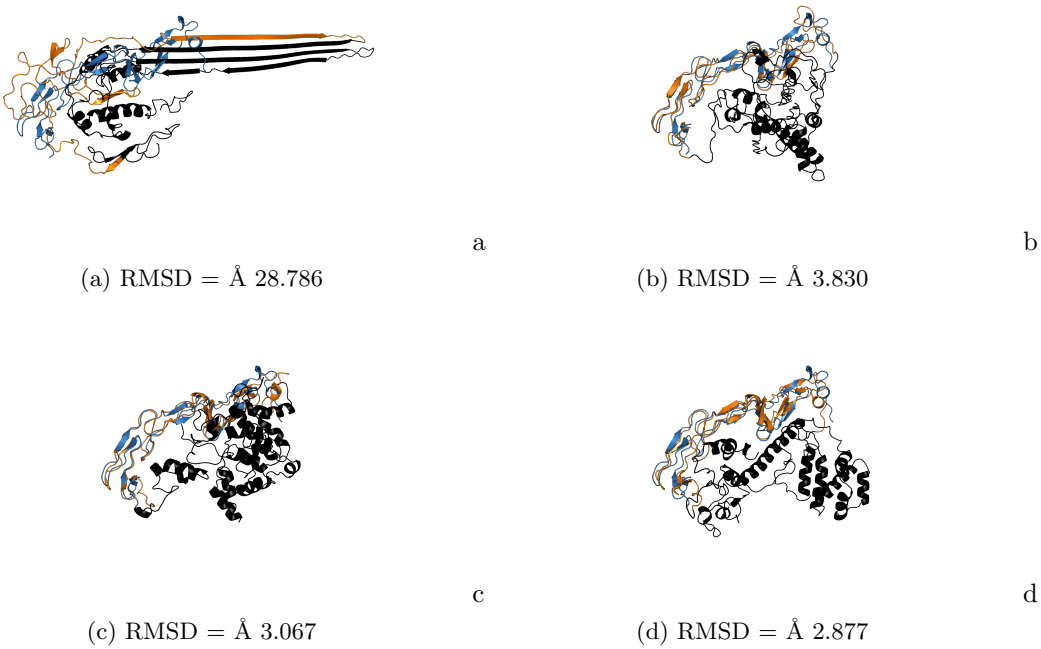


Figure 1: Both flowcharts illustrate the VIPUR pipeline wherein each block is a procedure the central circle is the purpose of the mutated applications and each arrow represents the path to it. Figure 1a has red arrows that indicate that both methods were incapable to produce the mutated PDB files. Within figure 1b the alternative method is proposed wherein PyMOL and PyRosetta (Sections 4.15, 4.4) is substituted by Modeller (Section 4.9) to acquire the mutated protein structures.

5.1.3 Expanding the VIPUR training set with data from TNFRSF1A by homology modeling and protein threading

Since the VTS did not have any features of TNFRSF1A (Section 3.2) the amino acid sequence was collected from Uniprot (Section 4.14) and the protein from RCSB (Section 4.13). The structures available of TNFRSF1A were incomplete, fragments for the TNF α and β binding site [] were available and its death domain that interacts with TRADD [] which plays a role in apoptosis (Section 3.2). To acquire a monomeric structure of TNFRSF1A two ab initio modeling web services I-TASSER and Robetta (Sections 4.8, 4.7) had been employed. Both were given the task to model the whole protein with and without a template to determine how well they could model a known structure and what it would form. Determination of the best model was based on the smallest root mean square deviation distance (RMSD)in Å , between a produced model compared to the X-ray crystallographic model of the TNFRSF1A binding site.



a

(a) RMSD = Å 28.786

b

(b) RMSD = Å 3.830

c

(c) RMSD = Å 3.067

d

(d) RMSD = Å 2.877

Figure 2: 3D structures of TNFRSF1A (2a, 2b: I-TASSER, 2c, 2d: Robetta) without (left: 2a, 2c) and with templates (right: 2b, 2d). The sky blue colored structure in each figure is an X-ray crystallographic model (1EXT) of the binding site of TNFRSF1A and the orange structure is the representation of that identical fragment in the model made by the web services.

5.2 Analyses of proteins variants TNFRSF1A

5.2.1 Requirements for determining structural and binding effects of protein variants

Protein variants can be assessed from multiple perspectives and together they can form a holistic view on how a protein works and how mutations affect its workings. However adding perspectives to the protein assessment makes it complex and requires expertise to determine its validity and contribution, therefore the analysis has been limited to basic structural information and also make the assessment inline with the VIPUR methods.

Various proteins consist of multiple chains that can be identical or different depending on their function [] and should be taken into account when assessing protein variants since one residue might alter the binding between chains and might alter the proteins formation. Different molecules and atoms that do not make up a protein but play a role in a pathway and function (ligands and co-receptors) are able to affect a proteins shape and can behave differently when a residue is mutated.

A different aspect that can change with mutations is the alteration in motions between structures which can allow or disallow certain movements to occur and with inhibit processes.

5.2.2 Single protein variant analysis approach and its tools

Before introducing mutations into a protein structure it is helpful to know if a mutation has been observed to avoid allocating resources to something that does not occur. Therefore three tables with observed TNFRSF1A mutations (Sections 4.10, 4.11, 4.12) have been combined with an R script (Section 4.19) into a single table consisting of two columns. The first column (split into three columns 1) contains strings that describes the: original residue, position and where it mutates to, the second column describes whether a formed mutation is harmful, with most mutations the effects have not been identified yet.

Original residue	Position in the protein sequence	New residue	Classification
Cys	44	Tyr	PATHOGENIC
Thr	44	Pro	PATHOGENIC
Thr	44	Ser	PATHOGENIC

Table 1: The format wherein mutations were filtered from the GAVIN, GenomAD and Infevers tables (Sections 4.10, 4.11, 4.12), describe whether a structural mutation is harmful or not. For many mutations it is unknown and other classifications are available, to view the whole table visit the supplementary.

For assessing variants in TNFRSF1A a structural fragment was used that contained TNF β (1TNR) [] and was made homotrimeric with PyMOL (Section 4.15) which results in six chains that emulate a bound TNFRSF1A with TNF β. The first column of the mutation table did not contain sufficient information to apply mutations correctly and within the PDB different numbering is used than in the amino acid sequence. To bundle the information and make it usable for introducing mutations a Python script (Section 4.18) has been written that combines the mutation table, PDB chains and the correct position within the sequence into a type of table which has sufficient information to mutate structures.

Iteration number	Filename	Chain	Residue index in chain	New residue
34	1tnr3_TNFA	R	0	TYR
34	1tnr3_TNFA	T	0	TYR
34	1tnr3_TNFA	S	0	TYR
35	1tnr3_TNFA	R	0	PRO
35	1tnr3_TNFA	T	0	PRO
35	1tnr3_TNFA	S	0	PRO
36	1tnr3_TNFA	R	0	SER
36	1tnr3_TNFA	T	0	SER
36	1tnr3_TNFA	S	0	SER

Table 2: The format that describes the mutations that should be made by Modeller (Section 4.9), with specifications of the: model, file. chain, residue index and the new residue. The whole table for TNFA and TNFB are visible within the supplementary.

To introduce mutations within PDB structures a Python script (Section 4.18) was written which used the generated mutation table (Table: 2) and a matching PDB structure, from the table; it acquires an iteration number which specifies if a mutation has to be stored in a single file or across multiple files; the filename serves as key which determine the PDB that should be used; letters specify chains, numbers are indices within the chains and the last column states the three letter residue where it should mutate to. When a structure is read in through the Python bindings of Modeller (Section 4.9) all non standard atoms and molecules are removed because Rosetta (Section 4.3) is not able to deal with those atoms. Just before mutagenesis takes place missing atoms are added to the structure that were difficult to determine with experimental methods(Section 1.3). After the insertion of all mutations a last guess is made where disulfide bonds are added between cysteines. With this process variants can be easily made, however the mutations do not present a correct protein because the mutated residues can put the protein in a high energy state.

In the attempt to make mutated structures behave more natural two tools from the Rosetta software suite (Section 4.3) have been used to minimize energies within protein structures. With the backrub application (Section 4.3.4) 1000 altered backbone models have been produced each with 10000 Monte Carlo moves (Sections 2.2). Each model generated has a set of properties that describe the formed molecule

Relax **

We specifically chose to asses known mutations from infevers. ** Density goes beyond 1 ***

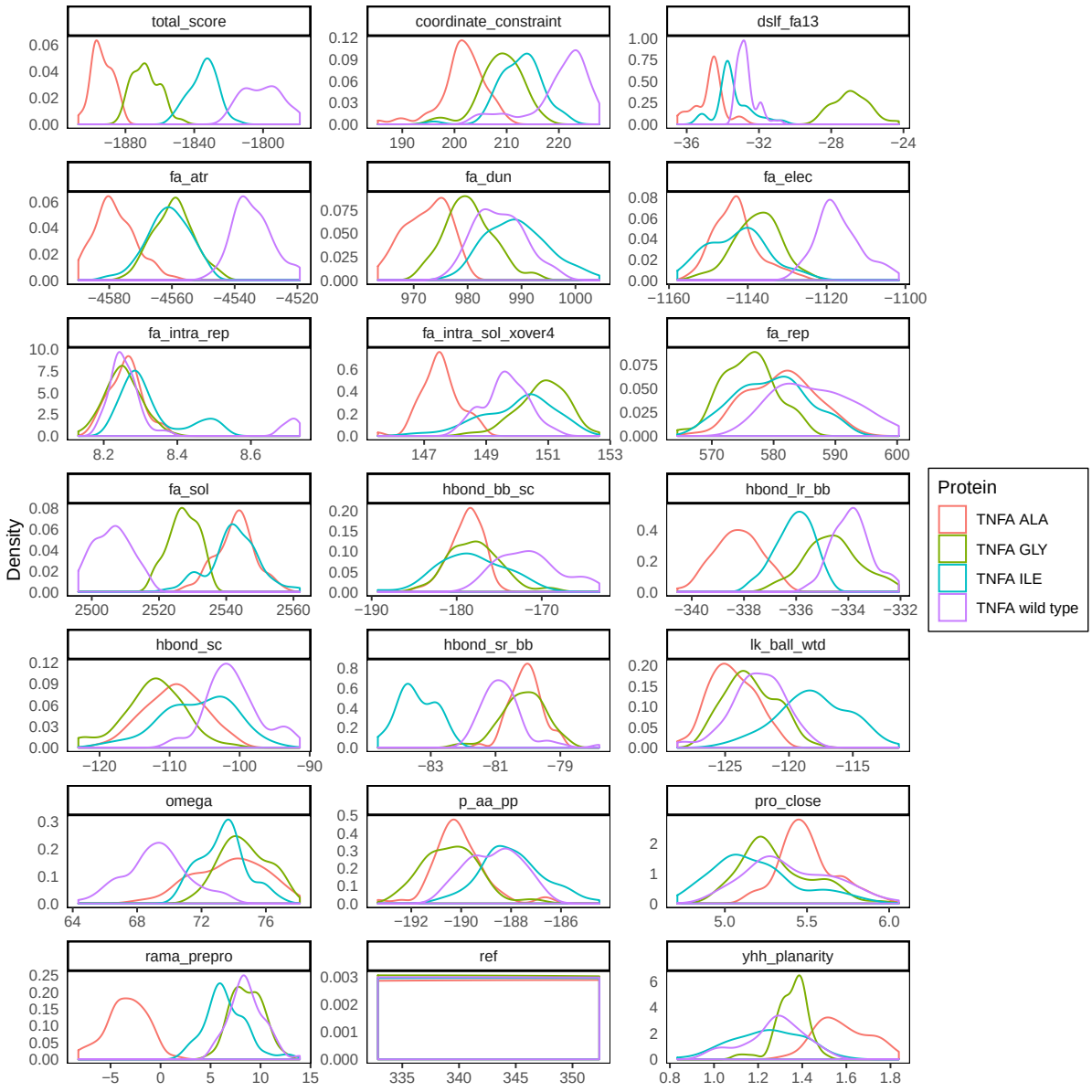

total_score
coordinate_constraint
dslf_fa13
fa_atr
fa_dun
fa_elec
fa_intra_rep
fa_intra_sol_xover4
fa_rep
fa_sol
hbond_bb_sc
hbond_lr_bb
hbond_sc
hbond_sr_bb
lk_ball_wtd
omega
p_aa_pp
pro_close
rama_prepro
ref
yhh_planarity
Density
Protein
TNFA ALA
TNFA GLY
TNFA ILE
TNFA wild type

Figure 3

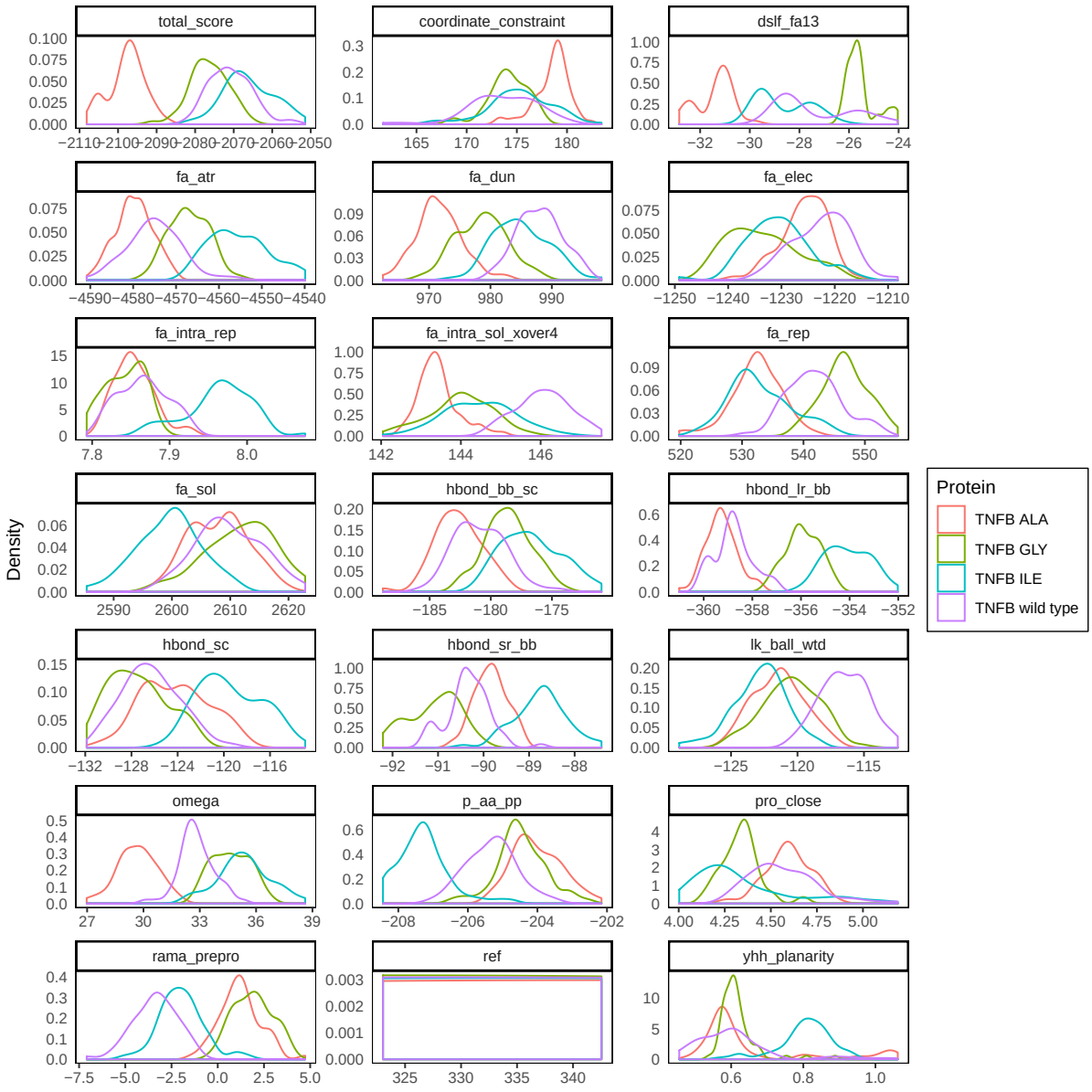

total_score
coordinate_constraint
dslf_fa13
fa_atr
fa_dun
fa_elec
fa_intra_rep
fa_intra_sol_xover4
fa_rep
fa_sol
hbond_bb_sc
hbond_lr_bb
hbond_sc
hbond_sr_bb
lk_ball_wtd
omega
p_aa_pp
pro_close
rama_prepro
ref
yhh_planarity
Density
Protein
TNFB ALA
TNFB GLY
TNFB ILE
TNFB wild type

Figure 4

5.3 Finding structural information and its mutations through HOPE

With uncertainty in the numbers provided by SPVAA a more textual informative approach was used with the web service HOPE (Section 4.16). The mutations that were known from Infevers (Section 4.12) and were also used with SPVAA were tested by HOPE (CYS62GLY, GLU138ALA and PHE141IlE), all reports are visible within the supplementary.

HOPEs first test was the mutation of Cysteine 62 to glycine, which is known within the Infevers table as pathogenic and was validated. It discovered that the residue was involved in a disulfide bridge and was 100% conserved in related protein sequences, based on the observation that cysteine formed a disulfide bridge it expected that with te replacement of it glycine would make the whole structures less rigid. HOPE predicts that mutation is pathogenic because of the high conservation of the residue, which is further confirmed by its search results in which if found the original publication where the observation has been described and associated to TRAPS [].

According to Infevers is the mutation of glutamic acid at position 138 mutated to alanine classified as likely benign and was not validated yet. HOPE discovered with a BLAST query that glutamic acid occurs often at position but other residues such as alanine have been observed at the position. Structurally glutamic acid forms salt bridges with proline 368 and leucine 390 and is found in a sequence of amino acids that is repeated through out TNFRSF1A. The amino acid lies within a domain where it interacts with other domains and is important for the proteins activity, with this mutation it might already perturb the binding capabilities according to HOPE.

The last mutation that was tested with HOPE was phenylalanine 141 to Isoleucine and was according to Infevers pathogenic and has been validated. Phenylalanine has been conserved at this positions and few other residues have been seen at the position, it is member of the identical domain as glutamic acid 138 and HOPE predicts that it would not damage the protein based on this information. However within the structure it could inhibit interaction with other domains and protein activity.

6 Discussion

People with rare diseases are currently hard to diagnose and are often not put in the desired treatment groups, with machine learning methods such as GAVIN 95% of the benign variants can be harmfully removed, however these methods rely on conservation and have been heavily optimized over the years []. A new refreshing approach called VIPUR uses sequential and structural data to predict deleteriousness of a protein variants.

Within the attempt to make VIPUR usable for the diagnosis of rare diseases it was discovered that some questionable steps were taken to make it especially applicable for diagnosis but also to determine deleteriousness of proteins itself: (i) "All protein models were standardized to remove unwanted components (duplicate chains, ligands, metals and non-standard amino acids)" []. Standardizing data can be beneficial to avoid learning features that do not matter and can reduce overfitting. However any form of context to the protein is removed and might therefore make incorrect assumptions about how: a monomer interacts with other monomers, ligands, metals, non-standard amino acids and water which can all have an effect on how proteins shape and interact []. (ii) With the utilization of Rosetta's Relax application different models are formed based on the Monte Carlo method (Section 2.2), VIPUR produces 50 structures with Relax per protein which is a tiny amount of the potential search space of possible folds that could have made changes in a mutated protein, which also can be seen in the scores of the model made from of TNFRSF1A with TNF α & β (Figures 3, 4), Rosetta itself suggests to make sufficient models, starting with 5000[]. (iii) The features acquired with probe in combination with the models that were produced, within the publication of Probe is mentioned: "It requires both highly accurate structures and also the explicit inclusion of all hydrogen atoms and their van der Waals interactions." []. It is currently not possible to determine if the structures were accurate, however is it likely that no loose hydrogen atoms were included within the structure based on the knowledge that all structures were standardized and likely some of the structures did not have any loose hydrogens within them. To make the outcome of Probe useful to VIPUR the program Reduce should have ran first, which add hydrogen atoms to the structure, which is recommend on the site were Probe can be downloaded from [].

**** easy pick **** Within Figure 4 of Robust classification of protein variation using structural modelling and large-scale data integration [] is stated that the mutation on position 204 from serine to proline is predicted deleteriousness, which is even without assessment highly likely because as can been in the structure it resides within an α helix and proline is known as the helix-stopper [].

Good other suggestions for finding if the approach really means something is by using shap [].

VIPUR uses PSI-blast to justify its results.

With the resource at our disposal we were unable to reproduce any of the results that were produced by VIPUR for testing purposes. The VIPUR pipeline could not be executed because it was not possible to compile PyRosetta or PyMOL on the cluster.

A good thing from VIPUR is that they took the information of all samples, it might introduce errors but also excludes lowers the extremes and might make it more realistic.

With Monte Carlo methods under te hood of the relax application it such as relax it is beneficial to produce many models since it is Although it is not part of developing a technique that can help diagnosing rare disease variant but the publication contains a claim ("VIPUR can be applied to mutations in any organism's proteome...." []) that likely contradicts with its methods ("remove unwanted components (duplicate chains, ligands, metals and non-standard amino acids)" []). There are 20 amino acids classified as standard, 22 are known to be incorporated into structures of some organisms [] and more variants are possible, which makes it unclear what is classified as a non-standard amino acid.

SPVAA method show similar weaknesses as VIPUR, it removes metal and water but keeps the ligands and are even added when necessary.

We did not asses the whole in a complex because there was insufficient information. (Membrane, water, other atoms and we could have left some atoms that were not water.)

With SPVAA only looked at the trimeric form an not the dimeric form of TNFRSF1A. Only using the Rosetta energy to determine the effect of the protein.

Isoforms were not taken into account.

Only assessed a small piece of TNFRSF1A and did not even look at the class of proteins itself.

With the resource at our disposal limited. More iterations on Rosetta relax than 64[].

The software written has a limited use currently and could be expanded to rapidly introduce mutations in multiple structures and chains at once.

It might have been useful to disable disulfide bridges if no cysteine residue is mutated because it is less likely that alterations are formed to disulfide bridges and otherwise they might be added without a reason.

Fragments were only available of 1EXT and 1TNR because it is a transmembrane protein which is difficult to make a structure from with X-ray crystallography. Maybe the first 30 residues were unnecessary because the might be signal peptides.

HOPE is informative with the results it produces, easy to use, fast and makes structural problems within proteins diagnoseable and understandable when a missense mutation is discovered. However it does not draws a solid conclusion and the information it collects depends on: previous publications, conservation and experimental structures, which will give it a disadvantage when limited knowledge is available. Also it does not asses a complex but it can describe binding sites from the monomer when previously discovered.

SPVAA's assement and VIPURs prediction could potentially be improved with the addition of molecular dynamic simulations to determine the effects of structural changes. With SPVAA it would have most likely become clear if the loss of the disulfide bridge from cysteine 62 in TNFRSF1A would have caused structural issues. VIPUR could benefit from it as new machine learning feature in situations where limited movement is observed and it changes tremendously when a missense mutation occurred in a protein increases with a mutation or vice versa.

All methods are reliant on existing structures making it difficult to asses it as a complex.

7 Conclusion

VIPUR is missing information to give a solid prediction about the deleteriousness of a protein variant and therefore likely suffers from errors in the predictions, SPVAA which uses changes in energy levels could be a more reliable source but also shows its weaknesses within the pace it can asses protein variants, it makes the determination manual and requires background knowledge for the assessment of protein variants.

of info that accurate we propsed another method for assessing protein structures within complex which may play a role in machine learning

Supplementary