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

May 30, 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 modelling 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
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
OS Operating System
PDB Protein Data Bank
PM Pivot Movement
PSI-BLAST Position Specific Iterative BLAST
PSSM Position Specific Scoring Matrix
RCSB Research Collaboratory for Structural Bioinformatics
RNA Ribonucleic acid
SASA Solvent Accessible Surface Area
SCOP The Structural Classification of Proteins
SLURM Simple Linux Resource Management
SODD SIlence 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

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 it is difficult to perform with some structures that contain flexible regions where in the positions of atoms is hard to determine the exact position [].

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 α and β (TNF) (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

VIPUR is a machine learning approach for predicting pathogenicity of proteins. The 106 features that were used for machine learning originate mainly (94%) from the Rosetta software suite (Section 4.3) applications; DDG monomer (Section 4.3.2), Relax (Section 4.3.1) and Rescore (Section 4.3.3), the remaining features were collected from PSI-BLAST (Section 4.5) and Probe (Section 4.6). All proteins in the VTS of which structures were known or had fragments available were collected from Modbase [] and SWISS-MODEL []. Proteins that did not have a structure within the databases were modeled with Modeller (Section 4.9 based on protein fragments that had the highest amino acid sequence identity to the protein. In some experimental determined structures duplicate chains, ligands, metals and non-standard amino acids were present, these inconsistencies are able to alter the features generated by software and could in some case hinder feature collection, therefor they were removed to make the data homogeneous. 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.

4.2 Simple protein variant analysis approach

Another approach for determining pathogenicity of a mutation is by assessing energy differences between a wild type and mutant protein residues inside its complex. Analyzing mutations from this perspective gives the ability to view a complex in whole and determine how residues cause perturbations in a complex. Missense mutations in monomers of complexes were made with Modeller (Section 4.9) and the backbone was refined with Rosetta's backrub application (Section 4.3.4), to lower the energy levels within side chains Rosetta relax (Section 4.3.1). This method shows similarities to that of VIPUR, was tested with TNFRSF1A (Section 3.2) and its ligands TNF α and β. This method keeps: duplicate chains ligands and metals within the structure, water is excluded since it can cause issues with Rosetta tools (Section **??**).

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 modeling []. Both approaches rely on the Relax (Section 4.3.1) for minimizing side chains. VIPUR uses rescore (Section 4.3.3) to acquire information about protein structures.

Both methods rely on Relax to minimize energies in the side chains of the remodeled structures. With DDG monomer both rely on energy minimization's in the side chains of the protein structures and need to information on energy changes in

The scores generated for the machine learning within the VIPUR approach rely on results generated by Rosetta software and to apply this approach the steps are reproduced. Several strategies were employed for realizing mutated structures, the first strategy was to identify the whole structure of proteins

The initial structure of the protein was produced with the application abinitio relax. For the prediction the application requires an amino acid sequence to identify homologous sequences in a curated database. Homologous sequences within the database are found by the BLAST algorithm, when a

For the search of the sequences it uses the BLAST algorithm and to find homologous amino acid sequences which have protein structures.

requires an amino acid sequence and it takes an amino acid sequence as input and searches in a curated protein database BLAST for finding homologous sequences.

to align sequences with to acquire homologous sequences. The homologous With these sequences it finds structures related to the protein For the prediction of the initial structure of TNFR the application abinitio relax was used.

With this tool a sequence is inserted as input that is aligned to

Missense mutated proteins have an altered amino acid that can cause differences in interactions with other amino acids, which can influence the backbone or side chain positions of a protein and therefore affect the structure. Software that makes missense mutations in protein structures (Modeller, PyMOL, PyRosetta) tend to replace residues without optimizing, causing odd energy levels or steric hindrance to arise.
Rosetta software suite Version 3.10

4.3.1 Relax

Relax was the only application used by both methods which tried to minimize energies in local conformational search space [] within the mutated structures. From each minimization attempt the structure was saved and scores for certain properties were calculated and written into a single file. From this score file VIPUR collected all samples and made 83 features out of it, the detailed approach used the scores from a single structure for its assessment.

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 4.2). 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 []. The used structures of TNFRSF1A (Section 3.3) were in complete and could therefore lack information regarding the structure when a mutation is introduced. To form a whole protein the, fragments of several known structures are joined by the Abinitio protocol within act on th with this web tool it was possible to predict the missing pieces of the protein

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 3.3) 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 3.3), 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. The table for TRAPS disease (Section 3.3) 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 4.3, 4.8) whole TNFRSF1A (Section 3.3) (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 and replacing TNF β with TNF α were done in PyMOL [], also PyMOL had some Python bindings to mutate proteins, which were used by VIPUR.
Version 2.2.3

4.16 HOPE

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 3.3) 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).
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

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) with the available classifications: Benign Pathogenic, Likely Benign, Likely Pathogenic, Population, Uncertain significance (VOUS) and Na.

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). The iteration number states if a mutation must be made in a single variant or in a different protein. Filename describes the protein to which the mutations are applied. Since structures can consist of multiple chains it has to be specified together with the index starting at 0 instead of 1 and finally to which residue it will be transformed.

6 Discussion

People with rare diseases are hard to diagnose

Prediction of pathogenicity in variants momentarily done based on sequence information and has been successful for certain groups of genes []. However pathogenticy of some genes with their variants cannot be classified by the currently used features for classification. Recently a method, called VIPUR, surfaced that incorporated sequence and protein data for classification of the pathogenicity from gene variants[].

In the attempt to reproduce the methods taken by the VIPUR approach on protein structures that are related to rare diseases it was realized that some questionable steps were being taken. With this approach all ligands were removed [] which changes the energies within and can therefor alter the structure [] and causes it to be analyzed from a single perspective instead of two when a bound ligand is also taken into account. Another step that was taken with VIPUR is that each structure is viewed as a monomer which is for some proteins not a problem, but for a complex that consists of multiple similar or a variety of different monomers makes it difficult to assess the effects.

To make predictions for new benign and pathogenic variants from TNFRSF1A, more information should be collected on how certain residues contribute to TNFRSF1A. More differences between interaction energies in mutated proteins could have been found be by adding molecular dynamics (MD) simulations of TNF α/β separately TNFRSF1A and combined with TNF docked into TNFRSF1A.

prediction of potential benign and pathogenic variants of TNFRSF1A isoforms should be included in the analysis to gain insight in which part of the proteins are highly important for the interactions and could result in a better prediction.

A significant contribution to gain more insight in how TNFRSF1A interacts with TNF α/β would have been the addition of molecular dynamic simulations; it shows how the proteins move on their own but also how the residues of the protein and the ligand interact with each other.

The VIPUR pipeline could not be executed because it was not possible to compile PyRosetta or PyMOL on the cluster.

however VIPUR has not been tested due to not having the correct software available and TNFRSF1A was not within the training data set of VIPUR.

Isoforms were not taken into account.

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." [].

The site of Probe mentions: "Meaningful analysis of molecular contact surfaces requires that ALL atoms are considered. Before using Probe, use the companion program Reduce to add hydrogens to the coordinate file." [], no mention of the Reduce software is mentioned in the VIPUR approach and therefor it is difficult to asses the meaning of previously acquired results in the VTS.

VIPUR is questionable because it has a limited amount of simulations. VIPUR uses PSI-blast to justify its results.

Describe the error made with apply mutations from different isoforms, see the table within results.

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

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

some steps have become questionable in structure of TNFRSF1A some questionable rare diseases some questionable training set some q to reproduce some of the results that were acquired with the VIPUR some questionable assu

Looking at the investigation t

With the resource at our disposal we were unable to reproduce any of the results that were produced by VIPUR for testing purposes, by

7 Conclusion

While VIPUR might be missing information to give a solid prediction about the pathogenicty of a protein variant, the detailed method used for determining the changes in energy levels could be a more reliable source for making predictions based of features.

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

Supplementary