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 10, 2019

Contents

List of Figures

Abbreviations

APAF1 Apoptotic Protease Activating Factor
DNA Deoxyribose Nucleic Acid
DISC Death-Inducing Signaling Complex
GAVIN Gene-Aware Variant INterpretation
MD Molecular Dynamics
TNF Tumor Necrosis Factor
TNFRSF1A Tumor Necrosis Factor Receptor Superfamily Member 1A
TRAPS Tumor Necrosis Factor Associated Receptor-Associated Periodic Syndrome
VIPUR Variant Interpretation Using Rosetta
VTS VIPUR Training Set

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 much for a quick and clear diagnosis. Current tools to predict variant pathogenicity rely on indirect evidence such as evolutionary conservation, annotation of regulatory genomics elements or structural DNA features. 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. However, this predictor was not integrated with the latest and greatest variant pathogenicity prediction approaches, was done on relatively small number of variants, and did not result in a tool that is ready to be taken into routine diagnostic practice.

1 Cell Death

1.1 Cell Death

Each human has about 37.2 trillion cells (3.72×10^{13}) [] of which several types are relative short lived [] compared to the life expectancy of a human in 2016 []. Continuously cells die by programmed cell death which is called apoptosis, this process allows to make certain features arise and keep cell growth in check []. The process of apoptosis can be triggered by pathways that activate caspases (proteases that cleave aspartate in proteins), once the process starts it is irreversible and the amount of caspases within the cell increases and is going disrupt the cells metabolism []. The internal system that determines when apoptosis initiates is the intrinsic pathway, it activates when their is internal stress in the cell such as damaged DNA or proteins (Which can caused by: heat, hypoxia, radiation, low/high ion concentration within a cell.) []. If stress is detected a mitochondrion releases cytochrome c into the cytosol and triggers a cascade, cytochrome c binds to apoptic protease activating factor 1 (APAF1) and starts to activate (intiator) caspase 9 that activates caspase 3 and thereby destroying proteins structures within the cell [].

are possible for activating the process and are caused by separate pathways. Both pathways lead to the activation of death-inducing signaling complex (DISC). This process is dependent on several

1.2 Tumor Necrosis Factor Receptor Associated Syndrome

1.3 Tumor Necrosis Factor Receptor Super Family Member 1A

1.4 Tumor Necrosis Factor Alpha and Beta

2 Protein Modeling Techniques

3 Monte Carlo

4 Materials and Methods

4.1 Two methods: scale and detail

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 4.2.5) applications; DDG monomer (Section 4.2.2), Relax (Section 4.2.1) and Rescore (Section 4.2.3), 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 4.7 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 4.14) by default or PyRosetta (Section 4.3) if PyMOL was not available.

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 4.7) and the backbone was refined with Rosetta's backrub application (Section 4.2.5), to lower the energy levels within side chains Rosetta relax (Section 4.2.1). This method shows similarities to that of VIPUR, was tested with TNFRSF1A (Section 1.3) 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 **??**).

4.2 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 4.2.1) for minimizing side chains and on DDG monomer (Section 4.2.2) to determine energy differences within the mutated protein. VIPUR uses rescore (Section 4.2.3) 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

4.2.1 Relax

4.2.2 DDG Monomer

4.2.3 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 4.2.1) and Backrub (Section 4.2.5).

4.2.4 Abinitio

4.2.5 Backrub

Missense mutated proteins might have altered backbone conformations depending on the mutation, mutations can result in energy differences among the structures and can therefor alter the whole structure.

differences in the backbone between a wild type protein and its mutated version can be done by detecting differences in energy within a proteins backbone.

Mutant proteins are able to have different backbone structures than a wild type protein.

Missense mutated proteins have an altered amino acid that can cause differences in interactions with other amino acids that can affect the backbone of a protein and therefore influence the structure. To discover such energy differences a Monte Carlo (Section 3) method

which can lead differences in energy by the contained positions

4.3 PyRosetta

4.4 BLAST

4.5 PSI-BLAST

4.6 Probe

4.7 Modeller

4.8 GenomAD

4.9 Infevers

4.10 RCSB

4.11 Uniprot

4.12 PDB

4.13 I-TASSER

4.14 PyMOL

Visualization of 3D structures, making images of proteins and putting the known orientations of monomeres in position were done in PyMOL []. Since some protein structures consist of multiple identical monomers they are left out of the structure and supplied with information about how the monomers are position to form the whole oligomer structure (Sections 4.10, 4.11).

5 Results

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.

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

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