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 17, 2019

Contents

List of Figures

Abbreviations

APAF1 Apoptotic Protease Activating Factor
API Application Programming Interface
Bash Bourne Again Shell
CPU Central Processing Unit
CSV Comma Seperated Values
DNA Deoxyribose Nucleic Acid
DISC Death-Inducing Signaling Complex
FEM Fixed End Move
GAVIN Gene-Aware Variant INterpretation
GRCh/hg Genome Reference Consortium Human Human Genome
MD Molecular Dynamics
MPI Message Parsing Interface
NCBI National Center for Biotechnology Information
OS Operating System
PDB Protein Data Bank
PM Pivot Movement
PSI-BLAST Position Specific Iterative BLAST
PSSM Position Specific Scoring Matrix
SLURM Simple Linux Resource Management
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 Variant Prediction In Genome Diagnostics

2 Protein Modeling Techniques

3 Monte Carlo

3.1 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.

3.2 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 [].

4 Cell Death

4.1 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

4.2 Tumor Necrosis Factor Receptor Associated Syndrome

4.3 Tumor Necrosis Factor Receptor Super Family Member 1A

4.4 Tumor Necrosis Factor Alpha and Beta

5 Materials and Methods

5.1 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 5.2) applications; DDG monomer (Section 5.2.2), Relax (Section 5.2.1) and Rescore (Section 5.2.3), the remaining features were collected from PSI-BLAST (Section 5.4) and Probe (Section 5.5). 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 5.11 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 5.18) by default or PyRosetta (Section 5.3) 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 5.11) and the backbone was refined with Rosetta's backrub application (Section 5.2.4), to lower the energy levels within side chains Rosetta relax (Section 5.2.1). This method shows similarities to that of VIPUR, was tested with TNFRSF1A (Section 4.3) 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 **??**).

5.2 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 5.2.1) for minimizing side chains. VIPUR uses rescore (Section 5.2.3) 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

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

5.2.1 Relax

5.2.2 DDG Monomer

The tool itself 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 [].

5.2.3 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 5.2.1).

5.2.4 Backrub

The backrub application is based on the Monte Carlo method (Section 3.2), 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 $\pm 10°$ []. It was used on the mutated protein to relax the modified backbone structure.

5.3 PyRosetta

Is an application programming (API) which has Python bindings (Section 5.10) for the Rosetta software suite (Section 5.2), it founds its use in VIPUR when no PyMOL (Section 5.18) was available to mutate residues within a structure [].
Version 4

5.4 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.

5.5 Probe

5.5.1 Robetta Protein prediction server

The web tool Robetta integrates several tools to form whole protein structures based on sequence alignments of previously discovered structures. It requires a protein a fasta file or 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 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 webtool it was possible to predict the missing pieces of the protein

5.6 I-TASSER

Predicts proteins with protein threading
Server version

5.7 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.

5.8 MPI

Some tools from the Rosetta software suite (Sections 5.2) 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

5.9 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 5.10) and submit jobs to the SLURM workload manager (Section 5.7).
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)

5.10 Python

Both VIPUR and the pipeline that minimizes backbone (Section 5.2.4) and side chain energies (Section 5.2.1) were written in Python due to its capabilities, ease of use and because modeller (Section 5.11) 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 5.13, 5.14, 5.12) with R Section 5.17) were filtered by a Python script.
Laptop version 2.7.15
Server version 2.7.11

5.11 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

5.12 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 4.4) with a missense mutation were filtered (Section 5.17) 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 [].

5.13 GenomAD

This 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 4.4), no classification was known for these mutations.

5.14 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 4.4) was used to collect missense mutations of TNFRSF1A gene.

5.15 RCSB

5.16 Uniprot

5.17 R scripting language

With R the tables from GenomeAD, GAVIN and Infevers (Sections 5.13 5.12 5.14) of TNFRSF1A missense mutations (Section 4.4) 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 5.10). *R scripting front-end version 3.5.2 (2018-12-20)*

5.18 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 5.15, 5.16).
Version 2.2.3

6 Results

The structures 1EXT [] and 1TNR [] (Sections 5.15 5.16) that represent TNFRSF1A (Section **??**) were incomplete and to fill in the missing pieces of the structure

7 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

8 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