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 6, 2019

Contents

List of Figures

Abbreviations

DNA Deoxyribose Nucleic Acid
MD Molecular Dynamics
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 x 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 [].

Other cells within a multi cellular organism are able to trigger the initiation of apoptosis with the extrinsic pathway. Once a cell excretes signal that it is not healthy Another pathway that can be initiate is cause by signal from different cells

which triggers the release of cythocrome c from the mitochondria that binds to apoptic protease activating factor 1 (APAF1) and activates the initiator caspase 9 . [].

A pathway that can controls apoptosis internally is called the intrinsic pathway and is triggered by internal stress such as damage to DNA and proteins which can be caused by: heat, hypoxia, radiation, low/high ion concentration within a cell, []

Within the process of apoptosis initiator caspases (proteases that cleave aspartate from proteins) are activated by factors that can trigger pathways that initiate apoptosis [].

Most cell types are capable of apoptosis when triggered by internal or external factors and known as apoptosis, to allow certain features to arise and keep cell growth in check. Within the process of apoptosis caspases , a family of proteases that focuses on cleaving aspartate in proteins [**?**], play a major role in breaking organelles and structures in a cell. Apoptosis

Apoptosis can be triggered by internal and external factors; or infection. [**?**]

Apoptosis can be triggered by internal and external factors;

is mainly triggered by two separate pathways that, the intrinsic pathway; which is a response internally from cell stress such as DNA damage or

When apoptosis is triggered it irreversible and amplifies its effect.

Death of cells is a continuous process that happens within every human. One form that is necessary to allow certain functions and to keep cell growth in check is apoptosis, this form of regulated cell death is triggered by

to behave properly within an individual is apoptosis which the cells within a multi cellular organism intact this form is regulated and called apoptosis. With

to occurs mainly in two forms; the unregulated form, which is defined as necrosis [**?**], which is generally caused by external factors that damage a cell and results in rupturing cells that spill organelles and cytoplasm. The regulated form is ,known as apoptosis [**?**] , which can be triggered by two known separate pathways both resulting in cell death. The intrinsic pathway is triggered by

or by receiving signals from other cells (extrinsic pathway) [].

billions die everyday by the mec apoptosis. Apoptosis is a regulated form of cell death which is induced by an unregulated manner is described as necrosis, parts of the cells can be missing or destory due to necrosis and apoptosis. already 50 to 70 billion (50-70 x ^{9}) die every day [] by the process of apoptosis.

T Cells within a multi cellular organism go through several stages depending on their function, but the last stage is always cell death.

This can have multiple causes, but a regulated and common form of programmed cell death is apoptosis and is highly important fo

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 TNFRSF1A

1.3 TNF alpha beta

1.4 TRAPS

2 Monte Carlo

3 Protein Modeling Techniques

4 Materials and Methods

4.1 Two methods: scale and detail

VIPUR is a relative large scale approach for analyzing protein structure variants and requires several tools from the Rosetta software suite (Section 4.2.4) to acquire the majority of its features, the remainders are collected from PSI-BLAST (Section 4.4) and the solvent accessible surface area (SASA) with Probe (Section 4.5). Features from the proteins acquired by analyzing the structures with Probe and Rosetta were collected from Modbase [] and SWISS-MODEL []. Not all structures that were used from the databases for machine learning were complete, there were structures of which only pieces were experimentally determined and for some no structure was available. To level this problem homology models were made with Modeller (Section 4.6 based on known homologous structures. In some experimental determined structures duplicate chains, ligands, metals and non-standard amino acids were present, these inconsistencies were able to alter the features generated by software and could in some case hinder feature collection, were therefore removed.

Another approach is by looking detailed individually at a single gene product, effects of mutations can be determined by replacing residues and remodeling the structure to determine the change. This approach has been done for TNFRSF1A (Section **??**)

With the VIPUR approach treating each protein similar by its current methods an alternative method the same by applying each method on all structures and a different method an uncertain how it will behave a more traditional approach was also take. A different approach that was taken by investigating a single protein (Section **??**) structure and itsthorough investigation of changes within the

Models within the VIPUR training set (VTS) had different Models acquired from t Proteins within the VIPUR training set (VTS) were a full structure had only parts or none Not every protein used within the VIPUR Training Set (VTS) had a whole structure or

Within the approach proteins that had no experimental structure available were homology modeled with Modeller (Section 4.6), parts and whole structures from proteins were collected from Another approac Another approach that was taken was to thoroughly investigate

Protein variants were analyzed in large groups by the VIPUR approach by several tools to which machine learning was applied.

4.2 Rosetta

For the prediction and analysis of protein structures the Rosetta software suite was used, it contains various tools for protein, antibody analysis and design []. 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

4.2.1 Relax

4.2.2 Abinitio

4.2.3 Backrub

4.2.4 Rescore

4.3 BLAST

4.4 PSI-BLAST

4.5 Probe

4.6 Modeller

4.7 RCSB

4.8 Uniprot

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

4.9 I-TASSER

4.10 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 4.7, 4.8).

5 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.

Isoforms were not taken into account.

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

6 Conclusion