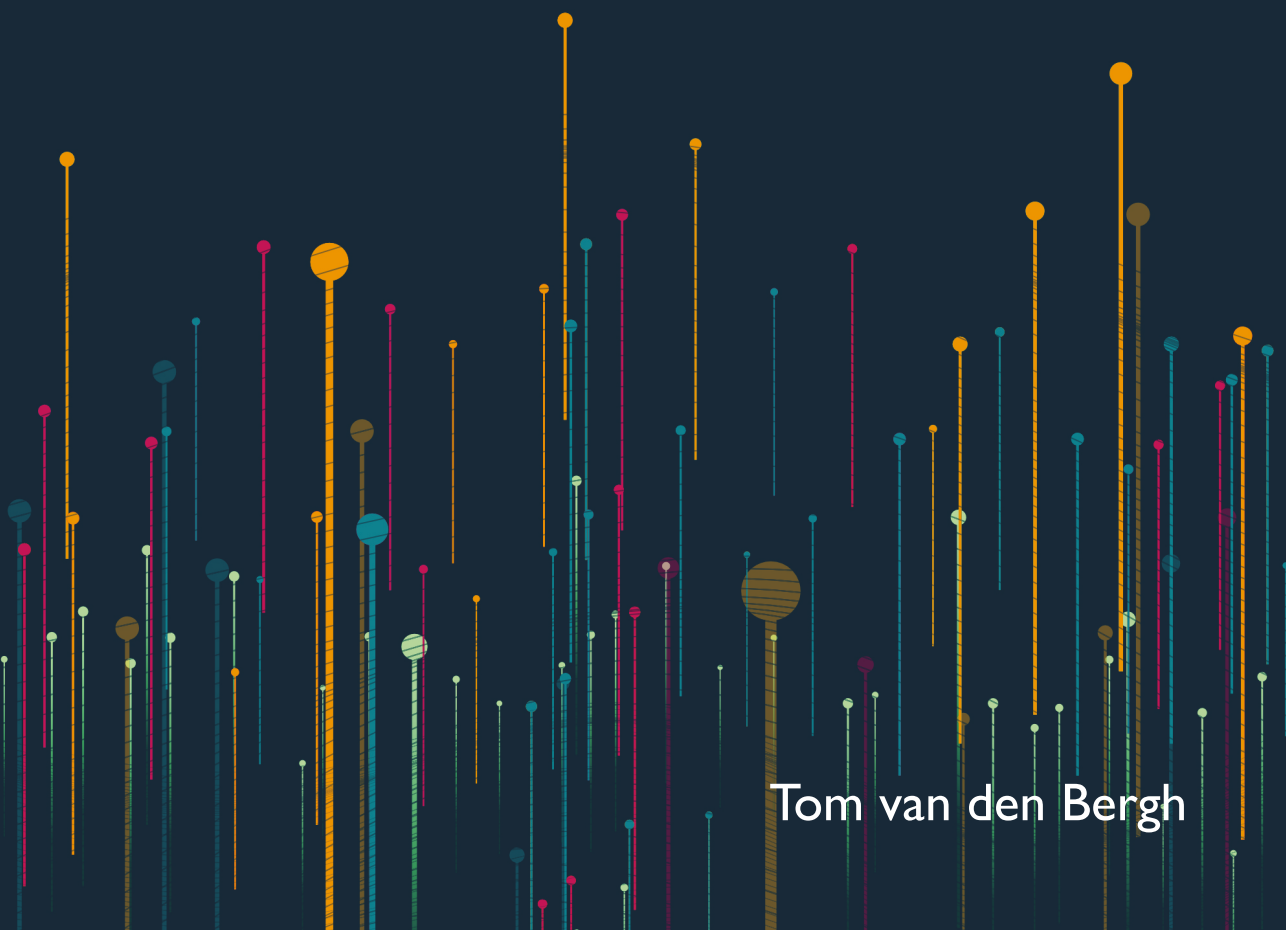
Towards a better understanding of proteins through **protein-family data integration.**



Tom van den Bergh

Propositions

1. Evolution is the ultimate wet-lab experiment.
   (this thesis)

2. When designing complex information systems,
   the real challenge is to keep it simple.
   (this thesis)

3. Phylogenetic trees are the proof of evolution.

4. Economic growth is not the only metric
   to measure the success of a nation.

5. Humanity is in a technological race against itself.

6. In the media, science should not be just another opinion.

Propositions belonging to the thesis, entitled

Towards a better understanding of proteins
through protein-family data integration

Tom van den Bergh
Wageningen, 2 February 2022

Towards a better understanding of proteins through protein-family data integration

Tom van den Bergh

Thesis committee

Promotors

Prof. Dr V.A.P. Martins dos Santos

Chairholder at the Laboratory of Systems and Synthetic Biology

Wageningen University & Research

Em. Prof. Dr G. Vriend

Emeritus professor of Bioinformatics of Macromolecular Structures

CMBI, Radboud University Nijmegen, Medical Centre

Other members

Prof. Dr M.H.M. Eppink, Wageningen University & Research

Prof. Dr M.W. Fraaije, University of Groningen

Dr A. Góra, Silesian University of Technology, Gliwice, Poland

Prof. Dr G.J. Poelarends, University of Groningen

This research was conducted under the auspices of the Graduate School VLAG (Advanced studies in Food Technology, Agrobiotechnology, Nutrition and Health Sciences).

Towards a better understanding of proteins through protein-family data integration

Tom van den Bergh

Thesis

Submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Wednesday 2 February 2022
at 4:00 p.m. in the Aula.

Tom van den Bergh
Towards a better understanding of proteins through protein-family data integration, 132 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2022)
With references, with summary in English

ISBN 978-94-6447-025-3
DOI https://doi.org/10.18174/557043

TABLE OF CONTENTS

CHAPTER 1

Introduction

1 PROTEINS

All living organisms produce proteins to perform the various tasks that allow them to survive and reproduce. Proteins are key to many different processes in and around living cells. They provide structure to the cell, are used by cells to communicate with their surroundings, store nutrients, catalyze reactions and can even protect the organism from harm. In short, without proteins a cell, and thus life, cannot exist.

1.1 Genetic information & protein synthesis

The collection of proteins that an organism produces is encoded in its genome in the form of deoxyribonucleic acid (DNA). The DNA of an organism contains all information needed for it to grow, develop and reproduce. A typical genome contains both coding as well as non-coding DNA. Large sections of a genome can be non-coding, for example, only less than 2% of the human genome consists of protein coding DNA. Genes, regions in the genome that encode functional molecules, need to be expressed to produce molecules such as functional ribonucleic acid (RNA) or proteins. Protein expression (illustrated in figure 1) is the process in which a ribonucleic acid (RNA) polymerase protein reads the DNA and transcribes it into messenger RNA (mRNA). mRNA can be translated by a ribosome protein into a chain of amino acids, or a 'polypeptide'. In this process each codon in a gene, consisting of three nucleotides, encodes for a single amino acid. The finished polypeptide folds into a 3-dimensional protein structure through various intrinsic and extrinsic forces such as hydrophobic interactions, hydrogen bonds and van der Waals forces. Many proteins are biologically active in this folded form. However, some proteins first need to be transported, modified or form a protein complex before they become active.

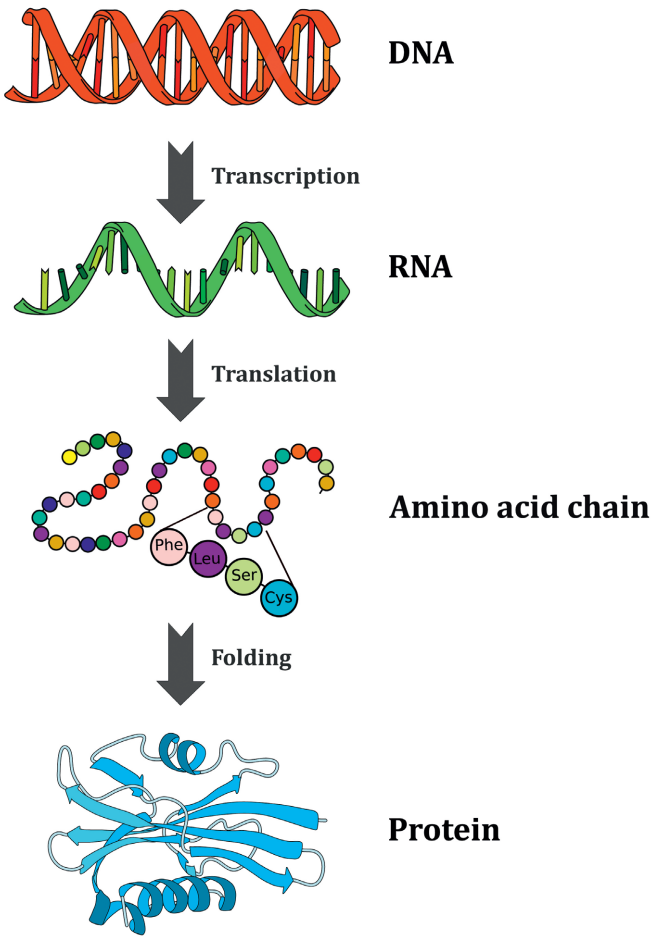


Figure 1. Schematic overview of protein expression. A gene in the DNA is transcribed to RNA molecules, that in turn is translated to a chain of amino acids. This chain then folds to form a 3-dimensional protein that can perform its function.

1.2 Protein structure

A protein can be described in four distinct aspects:

1. The primary structure, the sequence of amino acids that make up the polypeptide. This is the simplest description of a protein and commonly represented by a string of characters that each represents one of 20 amino acids used by the ribosomal translation machinery, 'ACDEFGHIKLMNPQRSTVWY'.
2. The secondary structure, the locally folded structures are subsegments of the complete conformation and form mainly by bonds with other amino acids of the protein molecule. Such structural regions consist of alpha helices or beta strands and are connected by unstructured loops of amino acids.

3. The tertiary structure, the completely folded 3-dimensional protein. The function of a protein results from the placement of key-residues in a specific 3-dimensional configuration. In this configuration each residue performs a specific task such as binding a compound or catalyzing a reaction.
4. The quaternary structure, this encompasses the mature protein in combination with other compounds and protein molecules needed to form a protein-complex and create a functional and biologically active molecule. Not all proteins require a quaternary structure.

While the average protein consists of around 350 amino acids, protein sizes can range from less than a hundred to thousands of amino acids. Titin for example, is a human muscle protein of almost 27 thousand amino acids in length. Such large proteins consist of multiple structural domains that vary in size and function to form a complex molecule adapted to specific needs of an organism.

Even though some proteins consist of long amino acid polypeptides, when compared to living cells they remain very small. The diameter of most protein molecules is smaller than 10 nm and whereas most cells can be seen with a microscope, protein molecules are too small to be observed using techniques that utilize visible light. To place protein molecules in perspective to the size of cells, a single *E. coli* cell can contain between four and a hundred million protein molecules while a larger human cell is estimated to contain one to ten billion proteins (Milo 2013).

1.3 Protein evolution

Evolution is a process that, over time, enables species to adapt to their environment. During reproduction, an organism's genome is replicated. Errors in this replication process lead to random changes in the DNA of each new generation. When such a random change invokes a missense mutation, it results in an amino acid change in the encoded protein. Such a residue substitution can result in a negative, positive or neutral effect on protein function. Mutations with a negative effect will often lead to a reduced fitness and result in an organism with a smaller chance to survive and reproduce. However, mutations with a beneficial effect can increase the chances of survival, and thus of reproduction.

Over time, this evolutionary pressure leads to the selection and propagation of beneficial mutations throughout a population while detrimental mutations are suppressed. Of course, when conditions in the environment change, other variations might suddenly benefit the organism's chance to survive, reproduce and enable a species to adapt.

Structure conservation

Evolutionary pressure counteracts the random changes in replication and leads to proteins that remain functional throughout generations.

This pressure applies solely on to protein function. A protein's sequence and 3-dimensional structure are free to change as long as the function is conserved. However, due to the close relation between protein sequence, structure and function the effects of evolutionary pressures can be observed in protein sequences and their structures. Duplication events, where an additional copy of a gene is created, can bypass this effect and allow one copy to mutate freely and potentially result in a protein with a novel function.

Since protein function is determined, to a great extent, by the 3-dimensional structure, proteins often remain functional when the structure of a mutated protein remains the same. In fact, many residues can mutate without greatly affecting the overall protein structure and its function. Aerts et al. demonstrated the robustness of proteins with the construction of several protein variants with 70 mutations that remained active (Aerts et al. 2013). The structural constraints related to function cause protein structures to remain conserved far longer throughout evolution than protein sequences (Illergård et al. 2009). The specific function of a protein is determined by a small subset of, usually conserved, key-residues positioned in a highly specific 3-dimensional configuration. When such residues mutate, the protein either becomes defective and is filtered out by evolutionary pressure or the protein function changes and serves the organism in a different task.

Snapshot of evolution

In evolution many protein variants have been and are continuously produced. Therefore, the sequencing projects undertaken in recent years can be considered as a snapshot of evolution. Due to evolutionary pressures the propagation of detrimental mutations has been suppressed so that most protein sequences observed today result from neutral or useful variations in evolution. Since this bias towards sequences from functional proteins is present in the sequence data that results from sequencing projects, many non-functional variations will be missing from this snapshot, as demonstrated by Jochens et al. (Jochens et al. 2010). The analysis of protein sequences across a family can therefore be employed to discover patterns and gain insight into protein residues. Protein positions can be identified that play certain key-roles in a protein and insight can be gained in residues that are conserved or absent in protein family members and when mutated will likely have a detrimental effect. To conclude, a lot can be learned by studying the evolution of protein families.

1.4 Human genetic variation & disease

Evolution has led to a wide variety of variants in human genomes. This process of mutation results in genetic variation among a population and enables the species to adapt and evolve. In 2017, over 324 million human variations were reported in dbSNP (Sherry et al. 2001; dbSNP 2017). Whereas most variants can be considered natural human variation, some variations are detrimental and lead to a disease phenotype. Almost 4,000 genes are reported in the Online Mendelian Inheritance in Man (OMIM) and

Human Gene Mutation Database (HGMD) databases that, when mutated, can lead to one of 6,000 monogenetic disorders (Hamosh et al. 2000; Cooper et al. 2010). However, compared to the 324 million human variations reported in dbSNP, only 142 thousand (<0.05%) nonsynonymous variations related to disease are reported in the HGMD.

The distinction between pathogenic or benign human genetic variation still remains difficult to predict (Mahmood et al. 2017). To predict the effect on a phenotype, multiple levels of complexity need to be considered. For example, when a variant protein has retained some of its function or a fallback mechanism exists, such as an alternative pathway that can take over a critical process in a cell, a disease phenotype might not establish (MacArthur et al. 2012). When human variants critical to the functioning of proteins are better understood, diseases could be recognized and treated earlier to perhaps even prevent a disease phenotype from establishing.

1.5 The need to better understand proteins

Proteins are complex molecules that vary greatly in form, size and function. They can be roughly divided into three groups; membrane proteins that are anchored in a cell's membrane, globular proteins that are often enzymes and can dissolve in water, and fibrous proteins that can provide structure to cells. Each protein has a specific role in the complex functioning of an organism and, when it malfunctions, can lead to disease. Proteins also provide opportunities for biotechnological processes and solutions. Therefore, both the fields of diagnosis/treating human diseases and the application of proteins in biotechnological engineering projects stand to benefit from a better understanding of proteins. Here, a subset of protein types is explored, along with examples of both human diseases they cause and their potential for biotechnological applications.

Receptor proteins

Receptor proteins enable a cell to react to signals. A signal can be passed along a signal transduction pathway that usually regulates gene expression. For example, an activated receptor can initiate cell reproduction, cell death or the production of a specific enzyme in response to a change in the cell's environment. Trans-membrane receptors, such as G protein-coupled receptors (GPCRs) are responsible for the transmission of regulatory signals between the in- and out-side of a cell while nuclear receptors can move around in the cytoplasm and initiate DNA transcription when activated. Mutations in receptor proteins can for example cause them to activate without the original signal. This can lead to various diseases such as carcinomas or diabetes (Kimple et al. 2014; Yu et al. 2018). As a biotechnological application, receptors can for instance be used to engineer bacteria to detect and report the presence of specific compounds, such as explosives (Shemer et al. 2015).

Enzymes

Enzymes, proteins that catalyze chemical reactions, are often globular in shape. These proteins can perform a wide range of reactions efficiently and are employed to help a cell regulate proteins, defend against threats, or process nutrition. Over 5,000 reactions are currently known to be catalyzed by enzymes (Schomburg et al. 2013). Some enzymes are only activated after they have been secreted to the extracellular environment, for instance to break down nutrients into digestible parts.

Many human disorders result from dysfunctional enzymes, for example an impaired alpha-galactosidase enzyme can lead to Fabry's disease, or an affected Janus kinase that is part of a signal transduction pathway can result in a cell proliferation signal that, in turn, can lead to cancer (Koulousios et al. 2017; Jolly and Van Loo 2018).

Enzymes discovered in nature can also be optimized or modified to produce compounds that are of industrial or economic benefit. To achieve this, a protein is optimized, cloned into a production strain and fermented. Protein molecules harvested from the fermented cells can be dried and used. This has been applied to a wide range of applications such as laundry detergents, biofuel production and the degradation of plastics (Wilson 2012; Harris et al. 2014; Vojcic et al. 2015; Austin et al. 2018).

Membrane transporters

Transporter proteins can facilitate the movement of ions and molecules through a cell's membrane. This can either be achieved by channels through a passive diffusion process or facilitated by carriers through active transport. Potassium channels allow the flow of potassium ions across the membrane to restore the membrane potential and allow cells to generate electrical signals. Mutations in potassium channels can lead to a variety of muscle and heart disorders such as long QT syndrome (LQTS). On the other hand, porin proteins are utilized in nanopore sequencing technologies (Deamer et al. 2016).

Structural proteins

Structural proteins are often fibrous proteins that aggregate and form long protein filaments. Such molecules have a structural function and are more resistant to denature than globular proteins. They are for example used to create an extracellular matrix to which cells can attach or to produce other structural material such as nails, hair, or spider silk. Defects in structural proteins can lead to muscle and growth disorders such as Marfan syndrome where the extracellular matrix is impaired (Bolar et al. 2012). In biotechnology research, fibrous proteins are of interest to engineer new materials with specific properties (Yigit et al. 2016). Besides the protein types discussed here, many other protein types exist that can lead to disease or can be engineered such as antibodies, hormonal messengers, regulatory proteins or even toxins.

Challenges & Opportunities

It remains a challenge to accurately predict the effect of novel variations in (human) proteins, to distinguish between human natural variation and pathogenic mutations or to determine which mutations should introduced to optimize a protein. A better understanding of protein function would accelerate the adoption and application of proteins in production processes as commercial products and help to diagnose and treat human diseases.

2 WET LAB VERSUS BIOINFORMATICS TO LEARN ABOUT PROTEINS

2.1 Wet Lab

A traditional approach to gain insight in proteins is to study the effect of introduced alterations to a gene or organism. Observations of modified or absent proteins can be compared with wild-type proteins or strains to deduce the natural role of a protein or residue. Measurements from such wet-lab experiments are direct observations of a protein in action and do not rely on potentially incomplete models or incorrect assumptions. Therefore, protein data derived from experiments can be considered a high value resource. Many different experimental methods are available to research proteins, such as gene knockouts or knock-ins, amino acid substitutions, alanine scans or mutation libraries (Wertman et al. 1992; Wang et al. 2015; Celie et al. 2016). One of the traditional methods to investigate a protein's role is to create a knockout-variant and observe an organism without this protein. To gain insight into the roles of specific protein residues, single mutations can be made to observe how a protein's function is affected. Screening of such mutations can be performed on many different parameters such as activity, substrate specificity, enantio-selectivity or thermostability to learn how a variation affects different traits.

To identify important positions in a protein an alanine scan can be performed (Morrison and Weiss 2001). This method consists of consecutively substituting all residues of a protein with an alanine, a residue with a non-reactive sidechain, to observe how this substitution affects the protein. The assumption is that substitutions of important residues have a detrimental effect while mutations of less important residues have a neutral effect on protein function. This method can identify hotspots in a protein that can be subsequently investigated with further mutagenesis experiments.

A more advanced approach is to construct mutation libraries where thousands of variants are screened to observe new combinations of residues that remain functional, this knowledge can for example be employed to select enzymes more suitable for a specific reaction. However, as this approach can be very labor intensive, the size of a library is usually determined by the ability to screen variants.

2.2 Alignments & information transfer

Although wet-lab experiments can offer valuable insight, they can be expensive and require much time and effort. Therefore, it makes sense to use data that is publicly available before wet-lab experiments are considered or to optimize wet-lab experiments. Evolution can be viewed as a wet-lab experiment on a very large scale that can be studied by sequencing samples of DNA taken from nature. Since evolutionary pressure creates a bias towards beneficial mutations in proteins that survive and are sampled today, the sequences available in nature can be used to learn about proteins without any experiments in a laboratory. For example, insight into a specific protein can be gained by an in-silico comparison with homologous proteins. To compare and analyze homologous proteins a protein alignment is needed that reveals which residues inhabit equivalent positions and perform similar functions. In essence, a protein alignment is the comparison of their 3-dimensional structures. Therefore, the highest quality protein alignments result from a superposition of protein structures. However, since for most proteins structural data is not available, proteins can often only be aligned by a sequence alignment method. Many sequence alignment tools are available to align two or more sequences (Katoh et al. 2002; Larkin et al. 2007; Potter et al. 2018). An alignment contains information on changes that occurred in evolution such as mutation, insertion and/or deletion events. A multiple sequence alignment (MSA), such as shown in figure 2, provides insight in how proteins changed over time under evolutionary pressure.

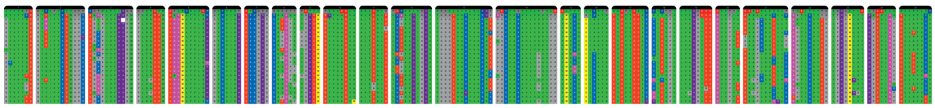


Figure 2. Part of a multiple sequence alignment. The residues are colored to indicate the chemical properties of the amino acids. Structurally non-conserved regions have been left out of this alignment.

Conservation, variation

An MSA can be analyzed to reveal patterns such as conservation and variation. On some alignment positions a residue will have remained conserved while on others a high degree of variation can be observed. Conservation often indicates that an evolutionary pressure is present to keep a certain function of these proteins intact. On the other hand, positions that show a high amount of amino acid variation indicate that mutations do not affect protein function too much. Such a basic analysis of evolution can already provide insight in important residues. For example, alignments of human protein sequences with only a relatively small number of close homologs have for years been used to evaluate pathogenicity of human natural variants (Ramensky et al. 2002; Ng and Henikoff 2003). However, MSAs that include more distantly related proteins contain more information and should thus be better suited to evaluate the importance of residues.

MSA information transfer

When experimental data is available for a closely related protein, a sequence alignment can be used to project information on a protein of interest. For example, when a protein structure is available, an alignment or a homology model can be created to learn which residues bind the ligand. This transfer of information between aligned residues is possible due to structural conservation of proteins in evolution. Since protein function is determined by the orientation of key-residues in the 3-dimensional structure, protein function changes when these key-residues mutate. However, due to the key position of these residues in the protein structure, substituted residues will likely perform a role similar to the residues they replaced. For example, when residues responsible for an enzyme's substrate specificity change, the specificity changes but the role played by introduced residues still determines substrate specificity. This allows for the transfer of residue specific information between structurally equivalent residues in homologous proteins (Gricman et al. 2014, 2015). Figure 3 shows an example of such equivalent residue positions for a protein family that includes distantly related proteins.

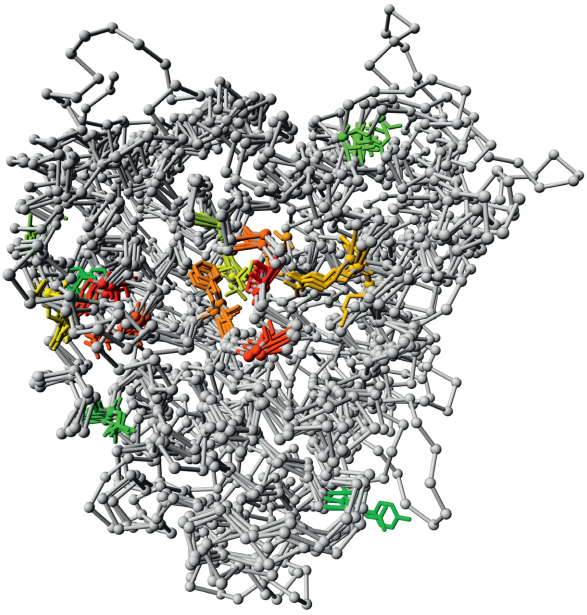


Figure 3. Superposition of five sequentially distinct P450 protein structures. The colored residues show a subset of positions that have remained structural equivalents throughout evolution and thus likely fulfill similar roles in these proteins.

Through insertion and deletion events the structural configuration of proteins can change throughout evolution. Some protein regions diverge so far that they no longer share a structural fold with family members. Such structurally non-conserved regions cannot be aligned to transfer information between residues of homologous proteins since they do not share structural equivalent residues. Especially in alignments of

protein families that cover large phylogenetic spaces (and large taxonomic ranges) such events must be considered.

Experimental data available for protein homologs can also be transferred via an alignment to further study a protein of interest. When data from remote members in larger protein family alignments are concerned, a high-quality MSA is needed to transfer information. A mistake in the alignment can result in an incorrect transfer of data to a residue that is not the structural (and thus functional) equivalent of the source residue.

To conclude, the study of protein evolution can provide insight by analyzing patterns in protein sequences sampled from nature. Knowledge gained from experiments on individual proteins can be combined with patterns observed in MSAs and projected on a protein of interest to better understand proteins.

2.3 Conclusions

Knowledge of homologous proteins can thus be a valuable resource to study a protein of interest. The combination of experimental results published in scientific journals with protein structures or sequences deposited in online databases allows for this information to become available to researchers before (further) wet-lab work is considered.

3 PROTEIN DATA AVAILABILITY

Over the past decades the amount of available protein related data has grown rapidly in all areas. Through these data researchers have observed many proteins in various environments to gain understanding in their diversity, form, and function. Different techniques have been developed to measure proteins and, therefore, the amount of data available in the form of protein sequences, structures and experimental results has expanded greatly.

3.1 Protein sequence data

The ability to determine a protein's primary sequence has enabled us to observe the results of evolution on a molecular scale. Protein sequences sampled from nature give insight in the diversity of proteins available in an environment and enable the discovery of new enzymes. To determine a protein sequence, the DNA that encodes it is typically isolated, sequenced and subsequently translated to obtain the amino acid sequence.

40 years of DNA sequencing

Over the past 40 years, many techniques have been developed that revolutionized the field of DNA sequencing (Shendure et al. 2017). In 1977, the first two techniques that allowed the identification of multiple nucleotides of a DNA strand were developed by Sanger et al. and Maxam and Gilbert (Maxam and Gilbert 1977; Sanger et al. 1977). Both methods created DNA fragments of different sizes where the last nucleotide was

known. By measuring the size of each fragment through gel electrophoresis the order of these fragments could be determined, thus revealing the nucleotide sequence. Sanger created these fragments by inhibiting the polymerase chain reaction (PCR) with modified nucleotides. Maxam and Gilbert created fragments by a chemical reaction that spliced the fragments at specific nucleotides. By 1987, a fluorescence-based approach of Sanger's method could detect around 1,000 base-pairs per day (Smith et al. 1986).

These sequencing techniques led to the development of shotgun sequencing. This method utilizes the overlap between sequence fragments sequenced from random clones to re-construct the DNA (Staden 1979). With the application of these techniques it became possible to sequence complete genomes (Fleischmann et al. 1995). These techniques also led to the human genome project (HGP) which took over 15 years and was finally completed in 2004 (International Human Genome Sequencing Consortium 2004).

NGS massively parallel sequencing

The next generation of sequencing technologies (NGS) became available in 2005 (Shendure et al. 2005). In NGS, many sequencing reactions are performed in parallel to greatly scale-up the sequencing rate, increasing throughput and dramatically lowering sequencing costs per base pair (Wetterstrand). These techniques enabled new large-scale experiments such as sequencing metagenomes of microbial communities and the re-sequencing of many whole human genomes (WGS) and exomes (WES). WGS and WES availability led to the 1000 genomes project and the release of over 6,500 human exomes (1000 Genomes Project Consortium et al. 2010, 2015; Fu et al. 2013). At around the same time, these technologies allowed expeditions over the world's oceans to be undertaken to sample and sequence microbial and viral metagenomes (Yooseph et al. 2007; Williamson et al. 2008). More recently, a worldwide ocean sampling day was introduced to survey ocean microbial communities over time (Kopf et al. 2015). These sequencing efforts allow protein families to be explored or human genetic variation to be charted.

Single molecule sequencing

Currently, two promising single molecule sequencing technologies are in development that could eventually replace NGS technologies. Single molecule sequencing is a new sequencing approach that no longer requires error prone DNA amplification steps and truly diverges from Sanger sequencing. These techniques work by either observing DNA synthesis in real time or by the detection of nucleotides that pass through a nanopore.

Pacific Biosciences uses an immobilized DNA polymerase that accepts nucleotides modified with different fluorescent tags. These tags emit light when the nucleotide is incorporated by the polymerase complex that can be detected to reveal the nucleotide (Eid et al. 2009). The sequential bursts of light reveal the nucleotide sequence and since the modified nucleotides do not terminate DNA synthesis, large read sizes can be

achieved. Due to a random distribution of sequencing errors, high-quality consensus sequences can be achieved through this technology and could offer advantages in the *de novo* sequencing of genomes.

Oxford Nanopore Technologies employs a nanopore protein in an artificial membrane to which a potential is applied that results in an electrical current flowing through the pore. When a strand of DNA passes through this nanopore the electrical current is influenced by the (combination of) nucleotides inside. These disruptions are detected and translated to reveal the DNA sequence (Deamer et al. 2016). This technology can lead to very large read lengths and, since no optical detection apparatus is required, can be reduced to small devices that can, for example, be used in the field with a laptop computer (Jain et al. 2018).

Sequence data

Due to the aforementioned development of sequencing technology, the amount of protein sequences available in public databases has grown exponentially over the past 20 years. Whereas the human genome project took almost 15 years to complete, it is now possible to sequence a human genome in a single day. Many sequencing studies have since been completed and this has culminated in the availability of over 100 million protein sequences in UniProt, even when protein sequences from closely related strains where removed in 2015 (figure 4).

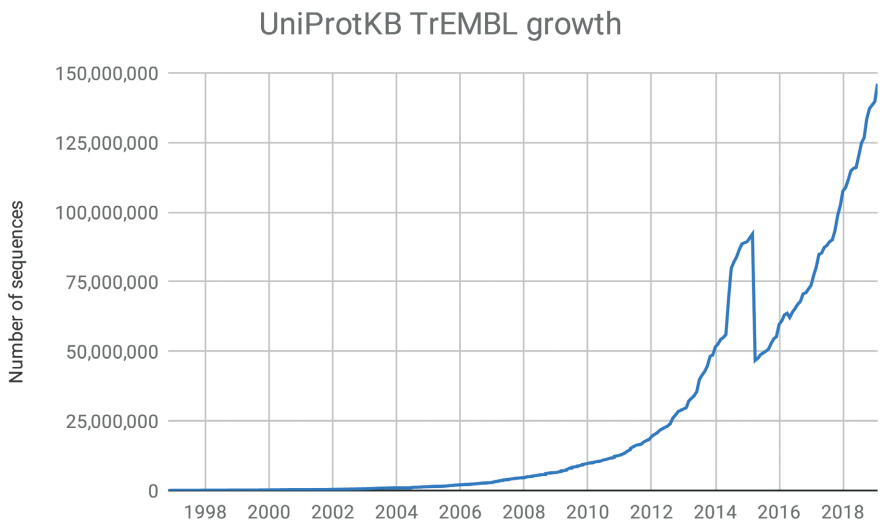


Figure 4. Number of protein sequence entries in the UniProtKB/TrEMBL database. The growth of sequences is exponential even after the removal of proteins from closely related strains in 2015, which nearly reduced the number of sequences by half. The drop of sequence entries resulted from a reduction in proteome redundancy.

3.2 Protein structures

Compared to an amino acid sequence, a protein structure can far better reveal how a protein functions on a molecular level. When available, a structure can provide insight into a protein's active site, surface area, interaction sites and other important positions. Therefore, the ability to elucidate the 3-dimensional conformation of a protein often advances the research in a field. However, the structure of a protein is far more difficult to determine than its sequence. Luckily, many protein structures are available in online resources such as the RCSB protein data bank (PDB) (Berman et al. 2000). Different methods are available to determine the structure of a protein. The three most popular methods are X-ray crystallography, nuclear magnetic resonance spectroscopy (NMR) and electron microscopy (EM). To date, over 130 thousand protein structures determined with one of these methods have been submitted to the PDB and are publicly available.

X-ray crystallography

X-ray crystallography is a technique that can determine a protein structure by the analysis of diffraction patterns generated by a beam of X-rays that pass through a crystalized protein sample. X-rays can interact with the electrons of a protein molecule and diffract into different directions. Unfortunately, these scattered X-rays cannot be focused with a lens to form a direct image of the protein molecule. However, since X-rays also behave as a wave, diffraction patterns can be observed and recorded. Since most X-rays pass straight through a molecule without interacting with any electron, a sample that contains a large number of proteins and a high intensity X-ray beam are needed to observe a diffraction pattern. A protein crystal is used, so that many protein molecules are positioned in an identical orientation and the diffraction pattern describes a single orientation of the molecule. Diffraction patterns from many different orientations of the crystal are analyzed to reconstruct an electron density model that ultimately reveals the protein structure. The main challenge with this technique is to obtain a protein crystal, especially for larger protein molecules. Usually, many different solution buffers will be attempted before a protein forms a crystal that is large enough to analyze. However, once a crystal is available the resolution that can be achieved is very high (<3 Angstrom). Therefore, around 90% of the structures present in the PDB are X-ray structures (Shoemaker and Ando 2018).

NMR spectroscopy

Nuclear magnetic resonance (NMR) spectroscopy uses a large magnetic force and electromagnetic radiation to influence the small magnetic fields that surround atomic nuclei that have spin. These nuclei absorb the emitted radiation at a certain frequency and resonate with the magnetic force for a short time at a high energy level. After a short time, the nucleus returns to a lower energy level and re-emits the absorbed radiation. This radiation is detected by the NMR device to create a frequency and intensity spectrum that describes atoms and their neighbors. By analyzing the different spectra, a 3-dimensional model of the protein molecule can be reconstructed. The resolution of

NMR structures is usually lower than that of X-ray structures. However, since it does not require a crystalized sample it is often used when a crystal structure is not available, as is often the case for trans-membrane proteins and large protein complexes (Wang et al. 2014). Around 9% of the structures present in the PDB are NMR structures.

Electron microscopy

Cryogenic electron microscopy (EM) of single protein molecules works by shooting a beam of electrons on a cryogenically frozen protein sample. Images that depict the sample from different angles are captured by the electron microscope. By combining the images from many different angles, a 3-dimensional reconstruction of the sample can be made. Traditionally EM images were recorded on silver films and were used to solve structures of large macromolecules with a relatively low resolution of at most 6 to 8 Angstrom (Vénien-Bryan et al. 2017). Since 2012, the development of direct electron detectors and improved image recognition algorithms have enabled electron microscopy techniques to solve structures smaller than the traditional 200-kDa barrier with higher resolutions that allow it to compete with X-ray crystallography or NMR techniques (Vénien-Bryan et al. 2017; Shoemaker and Ando 2018). In 2016, structures with a resolution below 4 Angstrom were solved (Merk et al. 2016) and recently, a hemoglobin structure was determined at 3.2 Angstrom (Khoshouei et al. 2017). Currently around 1.2% of the structures present in the PDB result from EM. However, since the molecular range and resolution that can be achieved with this method has started to match X-ray crystallography, this method is gaining popularity and has already surpassed the yearly number of NMR structures submitted to the PDB (Shoemaker and Ando 2018).

Structural classification of proteins (SCOP)

Proteins can be classified and clustered based on their sequence similarity. In this way the function of a protein can be inferred from sequential homologs. Although the accuracy of such function determination varies, it is a relatively simple approach and for example used in UniProt (Bateman et al. 2017). Alternatively, the 3-dimensional structures of proteins can be compared to infer a general function. Since a protein's structure remains conserved throughout evolution for a longer period than its sequence, a classification based on structural similarity allows for a clustering and function assignment based on more distantly related homologs (Illergård et al. 2009). The Structural Classification of Proteins (SCOP) is an online database that provides an index and classification of protein domains based on their structure (Fox et al. 2014). SCOP contains a hierarchical classification in which protein structures are assigned to a 'family' that contains close homologs, a 'superfamily' that also contains closely related families and a 'fold' that can contain several superfamilies. Structures are classified either manually or automatically. Version 2.07 of SCOP contains 1,232 folds, 2026 superfamilies and 4919 families (SCOPe statistics 2018).

3.3 Scientific literature

There is a huge amount of literature available to researchers on any single topic or protein family. The U.S. National Library of Medicine database PubMed currently contains over 28 million scientific publications. Last year this dataset expanded by 1.2 million papers and this number increases yearly. Due to the increase in publication rate, it has become nearly impossible for researchers to read all papers in their field. Often it might not even be possible to keep up with all papers on a single protein family that are published in a single year. This tremendous amount of available scientific publications requires methods that can automatically index reported experimental observations contained within papers by data mining and natural language processing. Without such methods, we risk that experimental results and insights being lost to future researchers by the sheer volume of publications.

4 PROTEIN FAMILY INFORMATION SYSTEMS

4.1 Integrate available data for a protein family

The huge amount of heterogeneous protein related data available requires the development of specialized storage systems to keep this data well organized. Due to the ongoing growth of generated data, it is increasingly difficult for researchers to optimally utilize all data available for a protein class. Molecular Class-Specific Information Systems (MCSIS) have been developed to store and integrate a wide variety of data available for protein families. Typically, such systems contain one or more alignments to connect homologous proteins and their residues and can be used to analyze sequences from the evolutionary perspective of a protein family. The integration of large amounts of heterogeneous data for a protein family allows insight into patterns across family members and can reveal molecular mechanisms of a protein class. Experimental results can be compared between family members or combined and projected on a single family member to gain a better understanding of key residues. Furthermore, experimental results can be placed in a protein family perspective and compared with patterns in the alignment to reveal the function behind observed evolutionary pressures. The ability to combine heterogeneous data for a complete protein family allows more insight to be gained than from the analysis of the individual proteins and can thus be considered greater than the sum of parts.

Alignments are typically stored in a database alongside protein (residue) annotations that are either manually or automatically collected. Due to the benefits that structured databases offer for the analysis of protein families, a number of MCSIS systems have been developed for various protein classes. The early databases concern protein classes that are of interest to the pharmaceutical industry. One of the first MCSIS systems was the GPCRdb, originally developed in 1998, for the analysis of G-protein coupled receptors (Horn et al. 1998). GPCRs, a class of receptor proteins involved in cell signal transduction, are related to many human disorders and form a major drug target worldwide.

Originally the GPCRdb contained hardly any structures and the important trans-membrane regions were aligned sequentially. Now complete protein structures have become available that can be aligned over larger regions of these receptors. The GPCRdb is still updated regularly and currently contains 270 protein structures and 15,090 sequences (Pándy-Szekeres et al. 2018). A recent example of how such a system can be used for pharmacogenomics is described by Hauser et al. (Hauser et al. 2018). The NucleaRDB, an MCSIS that comprises nuclear receptor proteins, was first developed in 2000 and contained 613 structures and 3,764 proteins at its last update (Vroling et al. 2012).

CYPED is an MCSIS for Cytochrome P450 monooxygenases that contains almost 600 structures and over 50,000 sequences (Sirim et al. 2009). These P450 proteins are of interest to both drug development and biotechnological applications. This database has been used to analyze, screen and engineer P450 proteins, for example by the comparison of different classes within this family or with text-mining approaches to identify specificity related positions (Gricman et al. 2014, 2015).

Other MCSIS databases include KLIFS: a structural kinase-ligand interaction database that contains almost 3,000 human and mouse protein kinase structures (Kooistra et al. 2016) and the abYsis MCSIS for antibodies (Swindells et al. 2017). Such an antibody system can be employed for the humanization of nonhuman antibodies that recognize antigens to reduce immunogenicity in human recipients of an antibody treatment.

3DM protein information systems are a type of commercially available MCSIS for protein superfamilies developed by Bio-Prodict. 3DM systems are generated by a standardized approach that can be applied to any protein family that comprises a conserved fold (Kuipers et al. 2010). Whereas KLIFS for example focuses on kinase-ligand interactions, 3DM builds on a structurally conserved 'core' alignment that allows users to interactively compare alignment position data between subsets of proteins or structures for a variety of purposes.

5 3DM SYSTEM GENERATION

The research described in this thesis is based on and utilizes 3DM protein information systems. The method to build 3DM systems has been improved over the past 10 years to become highly automated and only requires manual evaluation for quality control. The core of a 3DM system consists of a structure based multiple sequence alignment (SB-MSA) and a numbering scheme that unifies all residues. Via this numbering scheme each residue on a structurally equivalent position is assigned the same number. The process to generate a 3DM system alignments is described here.

To build a 3DM system for a target protein, first the superfamily of the structural domain of interest is identified using SCOP. The SCOP superfamily entry is parsed to extract

all protein family structures and a BLAST search (Altschul et al. 1990) is performed to identify homologous protein structures not present in SCOP. All identified protein structures are superimposed on a superfamily template structure that is closely related to the target protein. The software WHATIF is used to create this initial structural alignment (Vriend 1990). Transformation matrices for all superpositioned structures are stored for later visualization. From this initial structural alignment, sequentially distinct subfamily template structures are selected. These templates are typically selected to ensure their sequence identity with any other template is at most 80%. Therefore, each subfamily template represents a distinct phylogenetic segment of the overall superfamily. This approach significantly reduces the number of MSAs required while the phylogenetic space of the superfamily remains fully covered, as shown in figure 5. Moreover, structures elucidated in a similar configuration are preferred in this selection to optimize subsequent structural alignment accuracy.

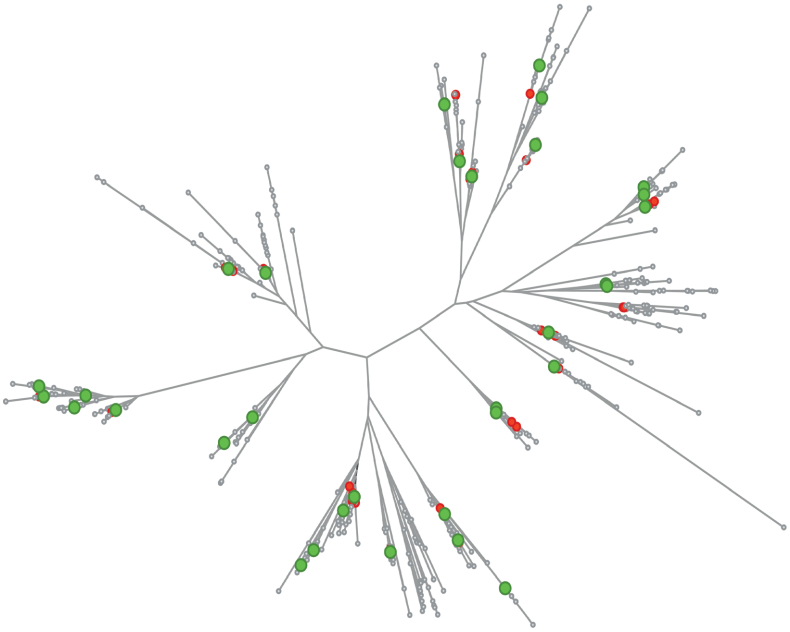


Figure 5. Phylogenetic tree of nuclear receptor superfamily with 519 protein structures (in red), alongside 500 protein sequences (in grey). From these protein structures 32 subfamily templates have been selected (in green). The protein sequences were selected from the superfamily alignment by a clustering approach to place the structures in context with the phylogenetic space of the aligned superfamily.

The subfamily representative structures identified in the previous step are used as structural alignment templates to iteratively superimpose all templates on each other. This results in a number of multiple structure alignments equal to the number of selected templates.

An autocompletion step then attempts to fill in any gaps in these alignments by comparing how structures are aligned in the different structural alignment results.

For example, to optimize the superposition of two structure we start by a direct method, comparing the superposition of protein A on protein B to that of B on A. When regions of structure A are absent from the alignment of A on B, but present in the alignment of B on A, then the alignment of structure A can be improved in the structural alignment of template B. This simple autocompletion is followed by an iterative, indirect approach.

The indirect comparison of structural alignments relies on the assumption that when protein A can be aligned with protein B and protein B can be aligned with protein C, protein A can be aligned with protein C as well. Using this assumption, structural alignments of remote homologue structures can be complemented.

To complement the alignment of structure A in the alignment result of template B, the alignment of A with B in the alignment result of template C can be evaluated. When residues of both structures can be superimposed on a region of structure C yet were absent from the alignment of A on B, the alignment result of template C is used to fill in this alignment. This indirect comparison of structure A with B this process is repeated with all alignment results other than those that used A and B as their super positioning templates. Furthermore, this autocompletion of A in the alignment result of template B can in later comparison rounds be used to complement alignments of other structures. This process is performed for every structure and iterated until the alignment can no longer be filled in. This method is comparable to a flexible structural alignment method that corrects for hinges in protein structures. Flexible alignment methods consistently outperform WHATIF in pair-wise structural alignments, however, due to a high number of false positive aligned residues, they result in protein family alignments with a low specificity. Therefore, we found that the high specificity of WHATIF structural alignments combined with this autocompletion method outcompeted flexible structural alignment methods in terms of specificity and accuracy of family structure alignments.

The consistency of the autocompleted alignments is checked to remove residues that are inconsistently aligned across different subfamily template alignments. The alignment with the highest total number of amino acids is selected from the autocompleted alignments and is divided into structurally conserved (core) and variable regions. Structurally conserved positions with a high number of template residues present in the structural alignment are allocated to the core. Insertions and deletions split up these core regions.

Structurally variable regions are kept out of the core (illustrated in figure 6C) since they did not remain structurally conserved throughout evolution. These residues don't have enough structural equivalents among the templates and thus, information cannot be

transferred between them. Optionally, the division of core and variable regions can be optimized to include more residues of a specific target protein. In that case, core regions present only in more closely related subfamilies are added to the core alignment.

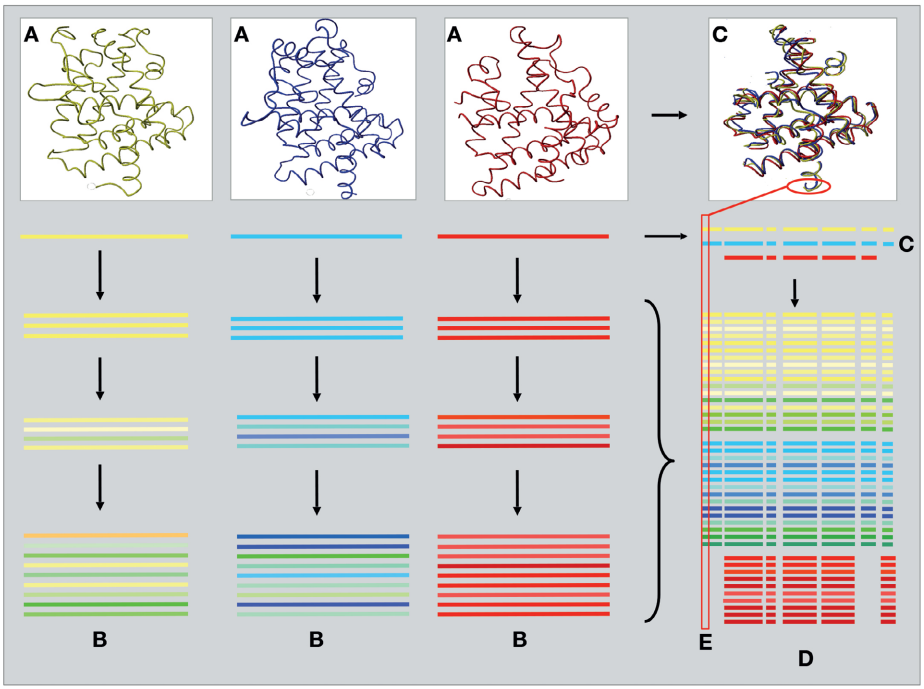


Figure 6. Visual representation of a structure based multiple sequence alignment. **A.** The three subfamily template structures. **B**. The iterative build-up of the subfamily MSAs. The colored lines represent distinct protein sequences. **C.** The core alignment shown in both a structure and sequence-like representation. **D.** The subfamily MSAs are combined into a structure based MSA that covers the complete superfamily. **E.** The column highlights the first residues in the structures and alignment. All such alignment positions are assigned a unique alignment number.

5.1 Structure based multiple sequence alignment

Each subfamily template is used as a seed to build a subfamily MSA. Related sequences are identified by a BLAST search against the UniProt, PDB and optionally GenBank protein sequence databases (Berman et al. 2000; Benson et al. 2013; Bateman et al. 2017). Structures that were left out of the subfamily template selection are included in this sequence alignment via the PDB BLAST search. The identified sequences are then aligned in an iterative profile-based alignment process that successively aligns increasingly more remote homologs in three or four rounds. The initial alignment profile is enriched with core/variable information for each template sequence position to prevent the placement of gaps and insertions in core regions. With each alignment iteration the subfamily alignment profile is updated with more distantly related sequences. This alignment process initially aligns sequences down to 40% identity over the core re-

gions (core identity) in three rounds. At this point, the various subfamily alignments are evaluated to determine positions that are conserved throughout the superfamily. This information is then used to update the subfamily alignment profiles to favor conserved residues at these core positions. The previous alignment steps are repeated with the addition of a fourth alignment round that includes sequences down to 30% core identity.

The structural alignment describes how subfamily templates should be aligned. This knowledge extends to the subfamily MSAs based on these templates. Therefore, the individual subfamily alignments can be combined into a structure based multiple sequence alignment (SB-MSA), as is illustrated in figure 6. This alignment consists only of core regions since the variable regions diverged structurally and can consequently no longer be aligned. Such alignments can expand to contain tens of thousands of sequences, dependent on the number of sequentially unique structural templates that are available for a superfamily.

An alignment numbering scheme is applied to all protein sequence and structure residues that make up the alignment so that residues on structurally equivalent positions are assigned the same number and can easily be identified. This alignment numbering scheme can be used to easily transfer data between all proteins in the superfamily alignment or visualize protein data on any of the integrated structures.

Quality control of sequence alignments

At low sequence similarities sequence alignments are prone to contain errors due to the limited amount of information present in protein sequences. However, high-quality MSAs are needed when alignments are used to transfer information between proteins. Therefore, the 3DM alignment method employs a unique quality control step. As most proteins are aligned in more than one subfamily and since the structural alignment contains information on how those subfamilies should align, the consistency of a protein aligned in multiple subfamilies can be verified. Specifically, for each core region it is checked that all relevant subfamily alignments aligned the same sequence fragment of a protein. When different sequence fragments of a protein are aligned to a structurally conserved region across multiple subfamilies, then at least one alignment is wrong. In such cases, this core of the protein can be excluded from the alignment. On the other hand, when different subfamily alignments consistently identify a sequence fragment for a core region, it is likely correct.

Collection of protein residue annotations

Once the alignment is complete, information is gathered for each protein residue and stored in the database connected to the alignment numbering scheme. To integrate protein related data, various data sources are used such as UniProt and KEGG (Kanehisa and Goto 2000; Bateman et al. 2017). All aligned protein structures are analyzed to identify residues that interact with ligands, DNA or other residues. Co-crystalized

compounds are recognized and integrated into the 3DM database alongside reaction data that is available for some proteins in the alignment. Protein mutation data are gathered from databases such as UniProt and dbSNP as well as from the biomedical literature. Scientific articles are mined to identify mutations, which are then linked to proteins and the alignment. This method is based on the MuteXt approach (Horn et al. 2004) and is described in more detail in chapter 2.

5.2 In silico analyses

A variety of *in silico* analyses can be performed on a superfamily alignment to investigate patterns left behind by evolutionary pressures. Residue conservation analysis can reveal important key-residues of a protein. When a residue is completely conserved in a protein superfamily, it indicates there was a strong evolutionary pressure to prevent mutations. Any mutations of such residues will likely result in a negative effect on protein function. Alternatively, when a residue is almost never found on a superfamily alignment position, the residue is likely not allowed at this position. This knowledge can be applied for the selection of residues for a "small but smart" mutation library (Jochens et al. 2010; Nobili et al. 2013) or consensus engineering (Aerts et al. 2013).

Correlated mutation analysis

Patterns that result from evolutionary selection can be observed in protein superfamily alignments. Evolutionary pressures preserve residues needed for protein function. When multiple residues are responsible for a function such as substrate specificity, ligand binding or protein dimerization, these residues are preserved together. Single mutations of such residues often do not result in viable residue combinations and are filtered out by evolutionary pressure. However, sometimes multiple mutation events take place over time that allow a protein to provide a (different) function. The result is that when we observe a sequence alignment, the non-active single mutants are absent since they have mostly been filtered out by evolution. As a result, it can appear that on certain pairs of alignment positions, residues change together throughout the alignment.

Correlated mutation analysis (CMA) can reveal the alignment positions where residues co-evolved throughout the evolution of a protein family (Kuipers et al. 2009). Since the cause of co-evolution is an evolutionary pressure, such residues often perform similar roles in a protein. However, since only a pattern in the alignment is observed, the function of the residues on these positions is not revealed by this analysis. Therefore, this function has to be investigated by other means, as shown in chapter 4.

Annotation hotspot analysis

One of the main advantages that an SB-MSA offers is that integrated amino acid annotations can be combined to identify hotspots, positions that can be mutated to modify or improve a certain function. For example, the number of mutations derived from literature or molecular contacts parsed from structures can be summed for each struc-

tural alignment position. This directly reveals important positions and can order these positions based on the number of annotations available for each position. Such insights often cannot be gained from the annotations of just a single protein.

The selection of annotations can be limited to take only those with a specific characteristic into account. For example, the indexed mutation data can be queried to retrieve only mutations that were mentioned in combination with a certain keyword, like specificity, or ligand contacts can be retrieved only from structures that contain an inhibiting compound.

When such annotations are grouped for each alignment position, they can grant insight into a much more specific characteristic than provided by a combination of all annotations. Moreover, a selection of alignment positions that were most often annotated could be used to start a mutagenesis experiment to utilize the observations of other experiments on homologous proteins. When positions that contact the ligand are analyzed with this approach, a distinction can be made between residues that contact the ligand in all members of the protein family and are thus important, and residues that only contact a ligand in a subset of protein structures available in this family. The same can of course be done for residues that contact a substrate, DNA molecule or another protein.

5.3 Usage of 3DM

3DM superfamily information systems have been generated for a wide variety of protein classes including GPCR's, P450's, α/β-hydrolases and many other enzymes (Kourist et al. 2010; Joosten et al. 2011). The tools in 3DM described above have enabled researchers to identify, investigate and improve proteins in these families. However, many of these are projects in the biotech industry and thus remain unpublished. Here, a few examples are shown that precede the work of this thesis and describe protein investigation, identification and engineering strategies with 3DM:

Joosten et al. identified a specific serine residue in oxaloacetate acetylhydrolases (OAH) that is essential for the reaction they catalyze; the hydrolysis of oxaloacetate to oxalic acid and acetate (Joosten et al. 2008). With the OAH 3DM system they were able to investigate sequences from fungi known to produce oxalic acid to identify this key serine residue. With this key residue they were then able to identify the previously uncharacterized oxalic acid producing enzyme in *Aspergillus niger*.

The “small but smart” mutation library design method mentioned above was tested by Jochens et al. and Nobili et al. to respectively improve the thermostability and enantioselectivity of an esterase (Jochens et al. 2010; Nobili et al. 2013). A 3DM system for the α/β-hydrolase fold enzyme family was used to select the residues to introduce at key positions. Only those residues that are common on these positions in the align-

ment were selected. A comparison with both a random (NNK) and inversed (uncommon residues) library showed that the number of functional mutants is much higher when only residues are introduced that are present in the alignment.

Kuipers et al. showed that with correlated mutation analysis protein positions can be identified that affect the specificity and activity of enzymes. Proteins from several families such as hexo-kinases, cupins and FAD-oxidases were mutated to confirm that such positions affected the specific-activity of these enzymes (Kuipers et al. 2009).

Likewise, the work described in this thesis not only entails hypothesis driven research but also facilitates more applied research by experts in the biochemistry of specific protein classes. Other examples of protein discovery, engineering projects and hotspot identification with 3DM and tools developed in this thesis are described in more detail in the following chapters and the discussion.

6 AIM & OUTLINE

The use of bioinformatics approaches to integrate and analyze protein related data for protein families deepens our understanding of protein function and offers distinct and novel opportunities for protein engineering and disease diagnostics. This work describes the developed strategies and opportunities to better understand protein function through the integration of protein data for complete protein families.

The chapters of this thesis describe different steps in an effort to collect, integrate and analyze protein related data for protein superfamilies (outline shown in figure 7). The first research chapter, **chapter 2**, describes the automated collection of mutational data reported in the literature for all proteins of the alpha-amylase protein superfamily and specifically for the human alpha-galactosidase that, when mutated, can lead to Fabry disease. This work shows that the extraction of such data from literature is a valuable data source and when integrated can lead to new insights. The amount of data available for a protein family would be tedious to collect manually and therefore, an automated approach was developed.

In **chapter 3**, we collect mutation data for two protein families and observe a significant overlap of alignment positions that contain mutations with a pathogenic effect in four very distantly related human proteins of the alpha-amylase superfamily. This implies that information from mutational data of one protein can be used to shed light on residues of other protein family members, even at large phylogenetic distances.

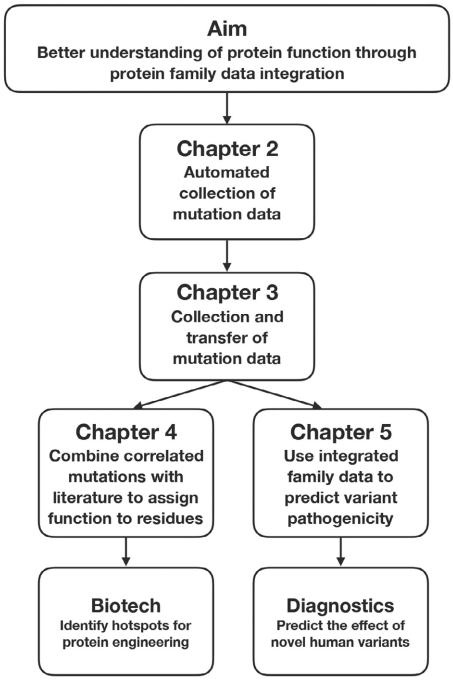


Figure 7. Graphical representation of the outline of this thesis.

Chapter 4 describes the use of mutation data mined from the literature to reveal functions that were preserved by evolutionary pressure. This evolutionary pressure resulted in groups of co-evolved residues that can be observed in an alignment. Certain residue positions of a protein family co-evolve throughout evolution and can be detected by correlated mutation analysis. However, this analysis does not reveal the (shared) function of residues on these positions. Therefore, we investigate if the combination of different data from a protein family information system could help to reveal the function of such residues and thereby help understand the biology of these proteins.

In **chapter 5** the data available in an integrated protein information system is used to train machine learning models that predict the pathogenicity of novel variants in human LQT-disease related proteins. This shows that the amount of data, and level of detail, can be used to describe amino acids in such a way that enables automatic approaches to evaluate the pathogenicity of novel variants. Combined, these chapters show an effort to collect, integrate and make use of protein data to allow a better understanding of proteins, their residues, and their functions.

Chapter 6, the general discussion, discusses the access, retrieval, quality and storage of the continuously increasing amounts of protein related data. We reflect on the applications described in the research chapters of this thesis and discuss new applications of integrated protein family information systems. Several examples are given of engineer-

ing projects that made use of 3DM systems in various ways, including tools developed in this thesis. These examples cover a selection of the ten other publications in which I collaborated and co-authored but that do not belong to the core of this thesis. These publications show that 3DM has been applied in projects to improve proteins, identify proteins with novel properties and taxonomic protein family studies. And finally, the discussion finishes with an outlook on the scale-up of 3DM from single protein families to a platform that covers the complete structural space. Such a platform, in turn, can be used for the generation of protein information systems for complete (human) exomes.

FUNDING

This research was co-funded by the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no: 289646 (KYROBIO). This research was co-funded by the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no: 289350 (NewProt). This research was co-funded by the European Union's Seventh Framework Programme (FP7/2013-2017) under grant agreement no: 613633 (SuSy). The work in this thesis was financially supported by the Dutch company Bio-Prodict B.V.

REFERENCES

1000 Genomes Project Consortium, Abecasis GR, Altshuler D, et al (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061–1073. https://doi.org/10.1038/nature09534

1000 Genomes Project Consortium, Auton A, Brooks LD, et al (2015) A global reference for human genetic variation. Nature 526:68–74. https://doi.org/10.1038/nature15393

Aerts D, Verhaeghe T, Joosten H-J, et al (2013) Consensus engineering of sucrose phosphorylase: the outcome reflects the sequence input. Biotechnol Bioeng 110:2563–2572. https://doi.org/10.1002/bit.24940

Altschul SF, Gish W, Miller W, et al (1990) Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Austin HP, Allen MD, Donohoe BS, et al (2018) Characterization and engineering of a plastic-degrading aromatic polyesterase. PNAS 201718804. https://doi.org/10.1073/pnas.1718804115

Bateman A, Martin MJ, O'Donovan C, et al (2017) UniProt: the universal protein knowledgebase. Nucleic Acids Res 45:D158–D169. https://doi.org/10.1093/nar/gkw1099

Benson DA, Cavanaugh M, Clark K, et al (2013) GenBank. Nucleic Acids Res 41:D36-42. https://doi.org/10.1093/nar/gks1195

Berman HM, Westbrook J, Feng Z, et al (2000) The Protein Data Bank. Nucleic Acids Res 28:235–242. https://doi.org/10.1093/nar/28.1.235

Bolar N, Van Laer L, Loeys BL (2012) Marfan syndrome: from gene to therapy. Curr Opin Pediatr 24:498–504. https://doi.org/10.1097/MOP.0b013e3283557d4c

Celie PH, Parret AH, Perrakis A (2016) Recombinant cloning strategies for protein expression. Curr Opin Struct Biol 38:145–154. https://doi.org/10.1016/j.sbi.2016.06.010

Cooper DN, Chen J-M, Ball EV, et al (2010) Genes, mutations, and human inherited disease at the dawn of the age of personalized genomics. Hum Mutat 31:631–655. https://doi.org/10.1002/humu.21260

dbSNP (2017) Database of Single Nucleotide Polymorphisms (dbSNP). Bethesda (MD): National Center for Biotechnology Information, National Library of Medicine. (dbSNP Build ID: 150). Available from: http://www.ncbi.nlm.nih.gov/SNP/

Deamer D, Akeson M, Branton D (2016) Three decades of nanopore sequencing. Nature Biotechnology 34:518–524. https://doi.org/10.1038/nbt.3423

Eid J, Fehr A, Gray J, et al (2009) Real-time DNA sequencing from single polymerase molecules. Science 323:133–138. https://doi.org/10.1126/science.1162986

Fleischmann RD, Adams MD, White O, et al (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269:496–512

Fox NK, Brenner SE, Chandonia J-M (2014) SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res 42:D304–D309. https://doi.org/10.1093/nar/gkt1240

Fu W, O'Connor TD, Jun G, et al (2013) Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493:216–220. https://doi.org/10.1038/nature11690

Gricman Ł, Vogel C, Pleiss J (2014) Conservation analysis of class-specific positions in cytochrome P450 monooxygenases: functional and structural relevance. Proteins 82:491–504. https://doi.org/10.1002/prot.24415

Gricman Ł, Vogel C, Pleiss J (2015) Identification of universal selectivity-determining positions in cytochrome P450 monooxygenases by systematic sequence-based literature mining. Proteins 83:1593–1603. https://doi.org/10.1002/prot.24840

Hamosh A, Scott AF, Amberger J, et al (2000) Online Mendelian Inheritance in Man (OMIM). Hum Mutat 15:57–61. https://doi.org/10.1002/(SICI)1098-1004(200001)15:1<57::AID-HUMU12>3.0.CO;2-G

Harris PV, Xu F, Kreel NE, et al (2014) New enzyme insights drive advances in commercial ethanol production. Curr Opin Chem Biol 19:162–170. https://doi.org/10.1016/j.cbpa.2014.02.015

Hauser AS, Chavali S, Masuho I, et al (2018) Pharmacogenomics of GPCR Drug Targets. Cell 172:41-54.e19. https://doi.org/10.1016/j.cell.2017.11.033

Horn F, Lau AL, Cohen FE (2004) Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors. Bioinformatics 20:557–568. https://doi.org/10.1093/bioinformatics/btg449

Horn F, Weare J, Beukers MW, et al (1998) GPCRDB: an information system for G protein-coupled receptors. Nucleic Acids Res 26:275–279

Illergård K, Ardell DH, Elofsson A (2009) Structure is three to ten times more conserved than sequence--a study of structural response in protein cores. Proteins 77:499–508. https://doi.org/10.1002/prot.22458

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431:931–945. https://doi.org/10.1038/nature03001

Jain M, Koren S, Miga KH, et al (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol 36:338–345. https://doi.org/10.1038/nbt.4060

Jochens H, Aerts D, Bornscheuer UT (2010) Thermostabilization of an esterase by alignment-guided focussed directed evolution. Protein Eng Des Sel 23:903–909. https://doi.org/10.1093/protein/gzq071

Jolly C, Van Loo P (2018) Timing somatic events in the evolution of cancer. Genome Biol 19:. https://doi.org/10.1186/s13059-018-1476-3

Joosten H-J, Han Y, Niu W, et al (2008) Identification of fungal oxaloacetate hydrolyase within the isocitrate lyase/PEP mutase enzyme superfamily using a sequence marker-based method. Proteins 70:157–166. https://doi.org/10.1002/prot.21622

Joosten, H.-J, Kuipers, et al (2011) 3DM Protein Engineering Super-Family systems applied to the P450 family

Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059–3066

Khoshouei M, Radjainia M, Baumeister W, Danev R (2017) Cryo-EM structure of haemoglobin at 3.2 Å determined with the Volta phase plate. Nat Commun 8:16099. https://doi.org/10.1038/ncomms16099

Kimple ME, Neuman JC, Linnemann AK, Casey PJ (2014) Inhibitory G proteins and their receptors: emerging therapeutic targets for obesity and diabetes. Exp Mol Med 46:e102. https://doi.org/10.1038/emm.2014.40

Kooistra AJ, Kanev GK, van Linden OPJ, et al (2016) KLIFS: a structural kinase-ligand interaction database. Nucleic Acids Res 44:D365–D371. https://doi.org/10.1093/nar/gkv1082

Kopf A, Bicak M, Kottmann R, et al (2015) The ocean sampling day consortium. Gigascience 4:27. https://doi.org/10.1186/s13742-015-0066-5

Koulousios K, Stylianou K, Pateinakis P, et al (2017) Fabry disease due to D313Y and novel GLA mutations. BMJ Open 7:e017098. https://doi.org/10.1136/bmjopen-2017-017098

Kourist R, Jochens H, Bartsch S, et al (2010) The alpha/beta-hydrolase fold 3DM database (ABHDB) as a tool for protein engineering. ChemBioChem 11:1635–1643. https://doi.org/10.1002/cbic.201000213

Kuipers RK, Joosten H-J, van Berkel WJH, et al (2010) 3DM: systematic analysis of heterogeneous superfamily data to discover protein functionalities. Proteins 78:2101–2113. https://doi.org/10.1002/prot.22725

Kuipers RKP, Joosten H-J, Verwiel E, et al (2009) Correlated mutation analyses on super-family alignments reveal functionally important residues. Proteins 76:608–616. https://doi.org/10.1002/prot.22374

Larkin MA, Blackshields G, Brown NP, et al (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948. https://doi.org/10.1093/bioinformatics/btm404

MacArthur DG, Balasubramanian S, Frankish A, et al (2012) A systematic survey of loss-of-function variants in human protein-coding genes. Science 335:823–828. https://doi.org/10.1126/science.1215040

Mahmood K, Jung C, Philip G, et al (2017) Variant effect prediction tools assessed using independent, functional assay-based datasets: implications for discovery and diagnostics. Hum Genomics 11:. https://doi.org/10.1186/s40246-017-0104-8

Maxam AM, Gilbert W (1977) A new method for sequencing DNA. Proc Natl Acad Sci USA 74:560–564

Merk A, Bartesaghi A, Banerjee S, et al (2016) Breaking Cryo-EM Resolution Barriers to Facilitate Drug Discovery. Cell 165:1698–1707. https://doi.org/10.1016/j.cell.2016.05.040

Milo R (2013) What is the total number of protein molecules per cell volume? A call to rethink some published values. Bioessays 35:1050–1055. https://doi.org/10.1002/bies.201300066

Morrison KL, Weiss GA (2001) Combinatorial alanine-scanning. Curr Opin Chem Biol 5:302–307

Ng PC, Henikoff S (2003) SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 31:3812–3814

Nobili A, Gall MG, Pavlidis IV, et al (2013) Use of "small but smart" libraries to enhance the enantioselectivity of an esterase from Bacillus stearothermophilus towards tetrahydrofuran-3-yl acetate. FEBS J 280:3084–3093. https://doi.org/10.1111/febs.12137

Pándy-Szekeres G, Munk C, Tsonkov TM, et al (2018) GPCRdb in 2018: adding GPCR structure models and ligands. Nucleic Acids Res 46:D440-D446. https://doi.org/10.1093/nar/gkx1109

Potter SC, Luciani A, Eddy SR, et al (2018) HMMER web server: 2018 update. Nucleic Acids Res 46:W200–W204. https://doi.org/10.1093/nar/gky448

Ramensky V, Bork P, Sunyaev S (2002) Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30:3894–3900

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74:5463–5467

Schomburg I, Chang A, Placzek S, et al (2013) BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. Nucleic Acids Res 41:D764-772. https://doi.org/10.1093/nar/gks1049

SCOPe statistics (2018) SCOPe 2.07: Structural Classification of Proteins. In: SCOPe Statistics. http://scop.berkeley.edu/statistics/ver=2.07. Accessed 4 Sep 2018

Shemer B, Palevsky N, Yagur-Kroll S, Belkin S (2015) Genetically engineered microorganisms for the detection of explosives' residues. Front Microbiol 6:. https://doi.org/10.3389/fmicb.2015.01175

Shendure J, Balasubramanian S, Church GM, et al (2017) DNA sequencing at 40: past, present and future. Nature 550:345–353. https://doi.org/10.1038/nature24286

Shendure J, Porreca GJ, Reppas NB, et al (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309:1728–1732. https://doi.org/10.1126/science.1117389

Sherry ST, Ward MH, Kholodov M, et al (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29:308–311

Shoemaker SC, Ando N (2018) X-rays in the Cryo-EM Era: Structural Biology's Dynamic Future. Biochemistry 57:277–285. https://doi.org/10.1021/acs.biochem.7b01031

Sirim D, Wagner F, Lisitsa A, Pleiss J (2009) The Cytochrome P450 Engineering Database: integration of biochemical properties. BMC Biochemistry 10:27. https://doi.org/10.1186/1471-2091-10-27

Smith LM, Sanders JZ, Kaiser RJ, et al (1986) Fluorescence detection in automated DNA sequence analysis. Nature 321:674–679. https://doi.org/10.1038/321674a0

Staden R (1979) A strategy of DNA sequencing employing computer programs. Nucleic Acids Res 6:2601–2610

Swindells MB, Porter CT, Couch M, et al (2017) abYsis: Integrated Antibody Sequence and Structure-Management, Analysis, and Prediction. J Mol Biol 429:356–364. https://doi.org/10.1016/j.jmb.2016.08.019

Vénien-Bryan C, Li Z, Vuillard L, Boutin JA (2017) Cryo-electron microscopy and X-ray crystallography: complementary approaches to structural biology and drug discovery. Acta Crystallogr F Struct Biol Commun 73:174–183. https://doi.org/10.1107/S2053230X17003740

Vojcic L, Pitzler C, Körfer G, et al (2015) Advances in protease engineering for laundry detergents. N Biotechnol 32:629–634. https://doi.org/10.1016/j.nbt.2014.12.010

Vriend G (1990) WHAT IF: a molecular modeling and drug design program. J Mol Graph 8:52–56, 29

Vroling B, Thorne D, McDermott P, et al (2012) NucleaRDB: information system for nuclear receptors. Nucleic Acids Res 40:D377–D380. https://doi.org/10.1093/nar/gkr960

Wang B, Li K, Wang A, et al (2015) Highly efficient CRISPR/HDR-mediated knock-in for mouse embryonic stem cells and zygotes. BioTechniques 59:201–202, 204, 206–208. https://doi.org/10.2144/000114339

Wang G, Zhang Z-T, Jiang B, et al (2014) Recent advances in protein NMR spectroscopy and their implications in protein therapeutics research. Anal Bioanal Chem 406:2279–2288. https://doi.org/10.1007/s00216-013-7518-5

Wertman KF, Drubin DG, Botstein D (1992) Systematic mutational analysis of the yeast ACT1 gene. Genetics 132:337–350

Wetterstrand K DNA Sequencing Costs: Data. In: DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). www.genome.gov/sequencingcostsdata. Accessed 26 Jul 2018

Williamson SJ, Rusch DB, Yooseph S, et al (2008) The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. PLoS ONE 3:e1456. https://doi.org/10.1371/journal.pone.0001456

Wilson DB (2012) Processive and nonprocessive cellulases for biofuel production--lessons from bacterial genomes and structural analysis. Appl Microbiol Biotechnol 93:497–502. https://doi.org/10.1007/s00253-011-3701-9

Yigit S, Dinjaski N, Kaplan DL (2016) Fibrous proteins: At the crossroads of genetic engineering and biotechnological applications. Biotechnology and Bioengineering 113:913–929. https://doi.org/10.1002/bit.25820

Yooseph S, Sutton G, Rusch DB, et al (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol 5:e16. https://doi.org/10.1371/journal.pbio.0050016

Yu S, Sun L, Jiao Y, Lee LTO (2018) The Role of G Protein-coupled Receptor Kinases in Cancer. Int J Biol Sci 14:189–203. https://doi.org/10.7150/ijbs.22896

CHAPTER 2

Novel tools for extraction and validation of disease related mutations applied to Fabry disease

Remko Kuipers*, Tom van den Bergh*, Henk-Jan Joosten, Ronald H. Lekanne dit Deprez, Marcel MAM Mannens and Peter J. Schaap

** These authors contributed equally to this work*

HUMAN MUTATION 2010, 2010 SEP;31(9):1026-32.

ABSTRACT

Genetic disorders are often caused by non-synonymous nucleotide changes in one or more genes associated with the disease. Specific amino acid changes, however, can lead to large variability of phenotypic expression. For many genetic disorders this results in an increasing amount of publications describing phenotype associated mutations in disorder-related genes. Keeping up with this stream of publications is essential for molecular diagnostics and translational research purposes but often impossible due to time constraints: there are simply too many articles to read. To help solve this problem, we have created Mutator an automated method to extract mutations from full text articles. Extracted mutations are cross-referenced to sequence data and a scoring method is applied to distinguish false-positives.

To analyze stored and new mutation data for their (potential) effect we have developed Validator, a web-based tool specifically designed for DNA-diagnostics. Fabry disease, a monogenetic gene disorder of the *GLA* gene was used as a test case. A structure-based sequence alignment of the alpha-amylase super-family was used to validate results. We have compared our data with existing Fabry mutation data-sets obtained from the HGMD and Swiss-Prot databases. Compared to these data sets Mutator extracted 30% additional mutations from the literature.

1 INTRODUCTION

Due to the ease of today's gene sequencing methods, the relation between genes and corresponding diseases has been unravelled for several genetic disorders. Moreover, the specific sequencing of disease-related genes in patients has enormously increased the available mutation data in the literature. For some extensively investigated genes, gene specific mutation databases are generated by extraction of mutational information from the literature. Examples of such mutation databases are the IARC TP53 Mutation database[1] and UMD p53 database for the tumor repressor gene TP53[2]. For molecular diagnostics and translational research these databases are used as reference to distinguish between naturally occurring SNPs and (potentially) pathogenic mutations in patients. Populating these databases usually requires manual intervention which makes it difficult to generate and maintain mutation databases. Therefore, up to date mutational databases are only available for a select number of disease-related genes.

In 2004, a tool MuteXt[3] was described for the automatic extraction of mutational information from literature. This tool was specifically designed for populating the nuclear receptor[4] and GPCR[5] Molecular Class-Specific Information Systems with mutation data. We have used the MuteXt method as basis for a new tool, Mutator, which can automatically extract and store mutational information from the literature for genes that are related to a genetic disorder.

Mutator was used to create a Fabry mutational database (FMDB). Fabry disease is an X-linked inborn error of glycosphingolipid catabolism that results from mutations in the alpha-galactosidase A (*GLA*; MIM# 300644) gene at Xq22.1. Currently two main Fabry disease related mutation data-sets exist; the Human Genome Mutation Database (HGMD)[6] and a collection of mutations automatically extracted from the UniProt databases[7]. The HGMD database is more complete since here mutational information is extracted from the literature. However, maintaining this database requires manual intervention. Our method extracts mutations from full text publications in a fully automated manner. The result shows an almost 100% coverage of mutations listed in the combined Uniprot and HGMD databases. Moreover, Mutator extracted from the literature 30% additional mutations covering 25% additional amino acid positions.

Human alpha-galactosidase is a member of the alpha-amylase protein super-family. In the past, it was shown that protein super-family derived data contextually stored in a Molecular Class-Specific Information System (MCSIS) can be used to describe individual functions of residues in proteins[8]. This has led to the development of the 3DM suite, a new generation MCSIS builder, that can semi automatically generate protein super-family systems specifically designed for mutant prediction purposes[9-12]. A 3DM super-family system is a knowledge base that contains and connects many different super-family related data types, such as structures, sequences, structure-based

multiple sequence alignments, protein-ligand interactions, mutational data, correlated mutation analysis results, and residue conservation. Mutator is part of the 3DM suite. The 3DM mutational data that is extracted from literature is collected by Mutator.

3DM was used to collect alpha-amylase super-family data and to generate the structure-based super-family alignment (3D-MSA). Strong correlations were observed between the aggregated mutational data and 3D-MSA derived data, which suggested that alignment derived data can principally be used to predict the pathogenicity of individual mutations in *GLA*.

On these principles Validator, a 3DM web-based graphical user interface, was developed for retrieval of literature extracted mutations and for validation of (new) amino acid variants (see supplementary figure S1). Validator uses various different information types, such as alignment information (e.g. amino acid conservation) and structural information (e.g. solvent accessibility, secondary structure information) that are stored in the 3DM database for variant validation. The predictability of each information type is pre-determined by examining how all known Fabry mutations relate to each specific information type. Furthermore, Validator generates a structure model for each variant in which bumps with neighboring amino acids are highlighted that are the result of the variant. These models can be viewed directly from the Validator website or can be downloaded, visualized and analyzed in the state of the art protein visualization tool YASARA. The newly developed Validator, the FMDB and the 3DM structure-based super-family alignment are freely available at http://3DMCSIS.systemsbiology.nl/FMDB/. The source codes of Validator and Mutator are currently an integral part of the 3DM commercial software suite. For other protein families commercial licenses can be obtained.

2 MATERIALS & METHODS

2.1 3DM Structure-based Super-Family Alignment Generation

The structure-based super-family alignment of the alpha-amylase protein super-family was generated as outlined by Folkertsma *et. al*.[13] and Joosten *et. al*.[9]. This method was automated in the 3DM suite, extensively reviewed by Kuipers *et. al*.[14] and is only briefly described here: All structure files from the SCOP[15] alpha-amylase family were extracted from the SCOP database to obtain a list of protein structure files with the alpha-amylase fold. The protein sequence of each distinct structure on this list was used as query to BLAST[16] against the PDB database[17] with a cut-off e-value of 0.005 to obtain a complete list of available structure files. For multi-domain proteins, only the alpha-amylase domain of the sequence was used as blast query to prevent inclusion of proteins that only contain a domain not related to the alpha-amylase super-family. Identical BLAST search settings were used for searches performed against the Swiss-Prot and TrEMBL[18] databases to collect sequences for which no structure

is available. 3DM was used to generate a structure-based super-family alignment from these sequences and structures in three steps:

1. The structure files were superimposed on the structure of the human GLA (pdb code 1R47[19]). From the resulting superpositioning, a structure-based multiple sequence alignment was extracted composed of structurally equivalent residues (core). Structural equivalence is defined as three or more consecutive residues that have their C-alphas within a 2.5Å sphere from the equivalent *GLA* residues.
2. The sequences of the resulting core alignment were divided into subgroups so that the sequences of each subgroup are no more than 80% identical to the next subgroup. For each subgroup a representative template structure is selected based on criteria such as the quality of the structure, the number of residues for which 3D coordinates are available in the structure, and the number of residues in the core as determined in step 1.
3. An iterative profile-based alignment procedure [20] (automated in 3DM) was used to separately generate subfamily alignments by aligning each super-family sequence to the most similar template structure. These separate subfamily alignments were combined to generate the ultimate super-family alignment using the core alignment from step 2 as a guide.

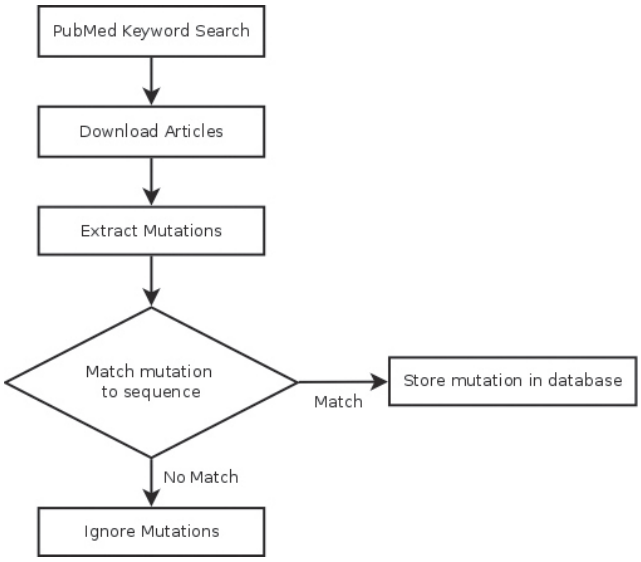


Figure 1. Schematic flowchart of Mutator program**.** See supplementary workflow S1 for the algorithm of Mutator and supplementary figure S2 for a more detailed flowchart of the algorithm.

2.2 Mutator

An overview of the workflow of Mutator is given in figure 1. To collect a large set of articles that potentially contain mutational information on *GLA* (or proteins homologues

to *GLA*), a keyword list was created. This list is used by Mutator to query the PubMed database to obtain a list of full text articles. Mutator collects mutations in four steps:

A) Retrieval of keyword selected (full text) publications; B) screening of the individual (full text) publications for mutational data using regular expressions; C) selection of sequences matching the wildtype subject protein sequence: D) overall scoring of combined feature of individual (full text) publications. For scoring of the mutations a Sequence Score (SQ-score) was used. Mutations extracted from publications that scored above the experimentally derived threshold levels were stored in a database. Details of the Mutator workflow (Fig. 1) are presented in supplementary figure S2 and supplementary workflow S1. Mutator was specifically designed to collect mutational information reported in proteins (or genes) related to certain diseases in patients. Therefore, in addition to the MuteXt method a module was added to Mutator that can detect mutations reported in DNA sequences.

2.3 Validator

Validator is a graphical user interface, specifically designed for DNA-diagnostic purposes. It can be used for variant analysis and retrieval of literature derived mutation data for a specific sequence of the 3DM database. After providing a mutation to the tool it returns the by Mutator extracted associated literature and a structural protein model visualizing the mutation including potential bumps with surrounding amino acids (Fig. 6). In addition, it predicts the likelihood that the mutation is pathogenic based on super-family alignment statistics such as (structural) conservation, amino acid distributions per alignment position (detailed in the results section). For a given mutation Validator also presents the Grantham distance[21], the Blosum62 substitution score[22], the solvent accessibility, and provides links to PolyPhen prediction tool [23] and the SIFT classification [24].

3 RESULTS & DISCUSSION

3.1 Mutator applied to Fabry disease: generation of the FMDB

This work presents a collection of *GLA* mutations retrieved from literature last accessed on 29 April 2009. It should be mentioned that the fully automated nature of Mutator enables continuous scanning of the literature and that the dataset presented here will soon be outdated.

Yip *et al* [7] recently described a method to retrieve single amino acid polymorphism data from the Swiss-Prot database. Specifically for *GLA* this dataset contains 137 mutations that cover 101 residues of the *GLA* sequence. The human gene mutation database (HGMD) contains mutations that are both automatically and manually collected. Excluding splice-site mutations, insertions, deletions, stop codons and frame shifts the free section of the 2009 version of the HGMD database contains 256 unique

point mutations that cover 166 residue positions of the *GLA* sequence. The restricted HGMD contains 301 unique *GLA* point mutations in total. HGMD describes a mutation only once. Mutator, however, stores references to all literature available of each specific mutation providing access to disease related metadata such as literature sources that contain variant phenotypic expression data in different patients. An overview of mutations available in the FMDB, HGMD and UniProt is available in supp. table S1.

Mutator uses a four step procedure to extract mutations from the literature; i) retrieval of keyword selected publications ii) screening of the individual publications for mutational data using regular expressions iii) evaluation of mutational data with respect to the corresponding subject sequence (here *GLA*) and iv) scoring of combined features above a set threshold. Supplementary table S2 shows the keyword list used by Mutator to query the PubMed database for Fabry disease related publications. This Fabry list resulted in the retrieval of 12,847 full text publications. From this set Mutator extracted and stored in the FMDB 1,781 mutations (371 unique mutations). All articles that Mutator selected for the first 100 *GLA* residues were manually examined for the presence of Fabry related mutational data. For these first 100 residues, Mutator collected 338 mutations from exactly 100 articles. Of these 100 articles, only six could be considered as false positives, since these six described mutations not related to Fabry. Three of these six articles described mutations in a human protein (human coagulation factor X) that contains a domain which is abbreviated with GLA. The other three articles described the human matrix GLA protein. To cope with this type of inconsistencies due to ambiguous keywords, extracted mutational data should also match the corresponding subject sequence. For example, when Mutator extracts the mutation G11V from a keyword selected text file, the program verifies that residue number 11 of the *GLA* subject protein sequence is indeed a glycine. In theory this step should reduce the false positive discovery rate for single extracted mutations to 5%. Besides having the right keywords all six articles contain mutational information at positions for which the *GLA* sequence has the same residue type such as the single G11V mutation described for the Gla domain of human coagulation factor X[25]. Therefore, an option was added that enables the user to provide a black list of keywords. Rerunning Mutator using “matrix gla protein”, “human factor” and “coagulation factor” as black list keywords removed these six false positive articles from the final set. Excluding splice-site mutations, insertions, deletions, stop codons and frame shifts, this set contains 1512 unique mutations and is highly (if not exclusively) populated with fabry related mutations. Comparison of this set of mutations with mutations stored in the HGMD and Swiss-Prot mutational databases showed that Mutator had collected 70 additional unique mutations. Six mutations were missed by Mutator because they were published in journals to which no subscription was available. This large set of fabry mutations enabled us to find correlations between pathogenicity of mutations in *GLA* and other data types that are stored in a 3DM database. Although here we have focused on *GLA*, it must be noted that the alpha-amylase super-family (see below) also includes human alpha-N-acetylga-

lactosaminidase (alpha-NAGA; alpha galactosidase B; MIM# 611458). Substitutions in alpha-NAGA can cause Schindlers disease. Upon switching from *GLA* to alpha-NAGA (Swiss-Prot: P17050) Mutator extracted from the literature all alpha-NAGA mutations that are reported in the OMIM database.

3.2 Alpha-amylase super-family alignment

The *GLA* gene is a member of the alpha-amylase protein super-family and protein super-family derived data can be used to describe individual functions of residues in proteins[13, 14]. The structures available of the alpha-amylase super-family can be divided into 41 sequentially distinct groups. The following structure files from the PDB database were chosen as representative structures to generate the super-family alignment: 1A47A, 1AMYA, 1AQHA, 1B2YA, 1BAGA, 1BF2A, 1BLIA, 1BVZA, 1EA9C, 1EH9A, 1G5AA, 1GCYA, 1GJUA, 1GVIA, 1H3GA, 1HVXA, 1IZJA, 1LWJA, 1M53A, 1M7XA, 1MXGA, 1QHOA, 1R47B, 1UD2A, 1UOKA, 1W9XA, 1WZAA, 2AAAA, 2BHUA, 2DH3A, 2E8YA, 2FH8A, 2GUYA, 2VUYA, 2Z1KA, 2ZE0A, 2ZICA, 3BC9A, 3CC1A, 3CZGA, and 3DHUA. The resulting super-family alignment contains 4,986 unique sequences and 217 structurally conserved positions (the core).

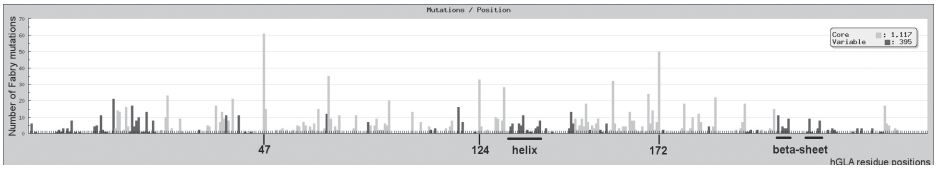


Figure 2. Number of independently reported Fabry disease related mutations per GLA residue position.

Independently reported mutations detected at structural conserved positions (core) are in light grey. Structural non-conserved positions are in dark grey. More than 40 independently reported mutations were extracted for 3D-positions 47, 124, and 172 corresponding with R112, N215, and R301, respectively of the *GLA* amino acid sequence. Note that, although only 50% of the *GLA* residues are core positions, the large majority of the mutations (1117) are observed at those positions.

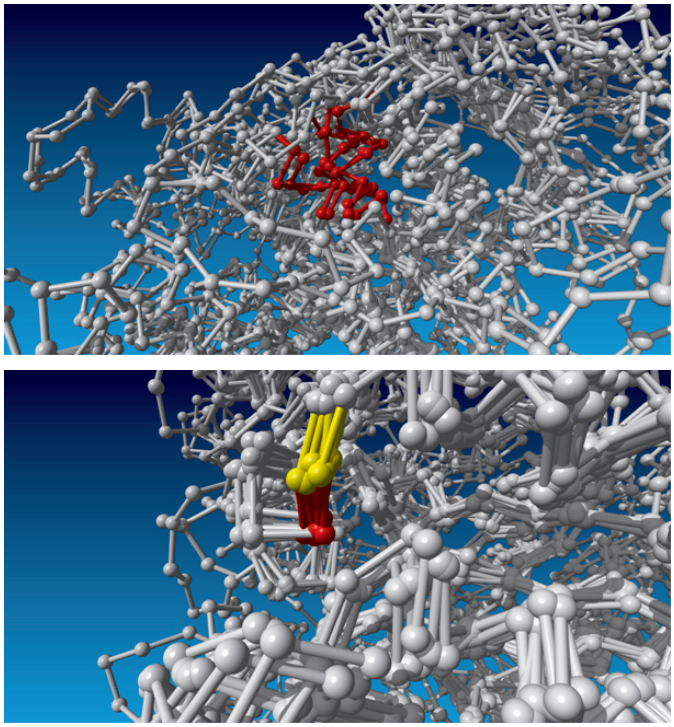


Figure 3. YASARA ball and stick backbone representation of seven superimposed protein structures of different sub-families of the alpha-amylase super-family.

3A: In red equivalent helices from the seven proteins. This helix is present in almost all super-family members but could not be included in the core due to variable positioning within the crystal structures. **3B:** The yellow and red colored residues are part of a structural highly conserved loop at 3D positions 47 and 48, respectively. 3D-postion 47 is a highly conserved glycine. 3D-postion 48 is in *GLA* a phenylalanine and the most reported mutated residue.

Figure 2 shows that 80% of the reported mutations are at structurally conserved positions (core). Two regions outside the core are highly populated with Fabry related mutational data. These two regions are a helix in the middle of the *GLA* sequence and a beta-sheet at the C-terminal end of the protein (Fig. 2) and contain 77 and 72 mutations, respectively. These two regions are present in most alpha-amylase structures. However, due to positional variability within the super-family structures the superposing of these regions was ambiguous (Fig. 3a). These two regions were therefore not included in the core. A more straightforward approach to determine structural important positions would be to assign structural importance only to residues of secondary structural elements (e.g. helices and beta-sheets). The advantage of such a method is that only the structure of the target protein (here *GLA*) is needed. However, it should be noted that, even though the core mostly consists of secondary structural elements, using only secondary structural elements as a delimiter is no solution. For instance, the residue position with the highest number of extracted Fabry related mutations (3D-number 47;

Fig 2) is not part of any secondary structural element, but is positioned in a structural highly-conserved loop (Fig. 3b) located at the outside of the protein. Additionally, alignment position 48 which is also part of this loop is a highly conserved glycine residue, which demonstrates that important residues are not exclusively located in secondary structural elements. If both core and secondary structural elements are considered to be structural important positions, 84% of all Fabry related mutations are linked to this group. This result suggests that it is 5 times more likely that a random mutation will result in manifestation of Fabry disease if this mutation involves a structural important position. The Validator tool (see below) therefore defines both core and secondary structural elements as structural important positions.

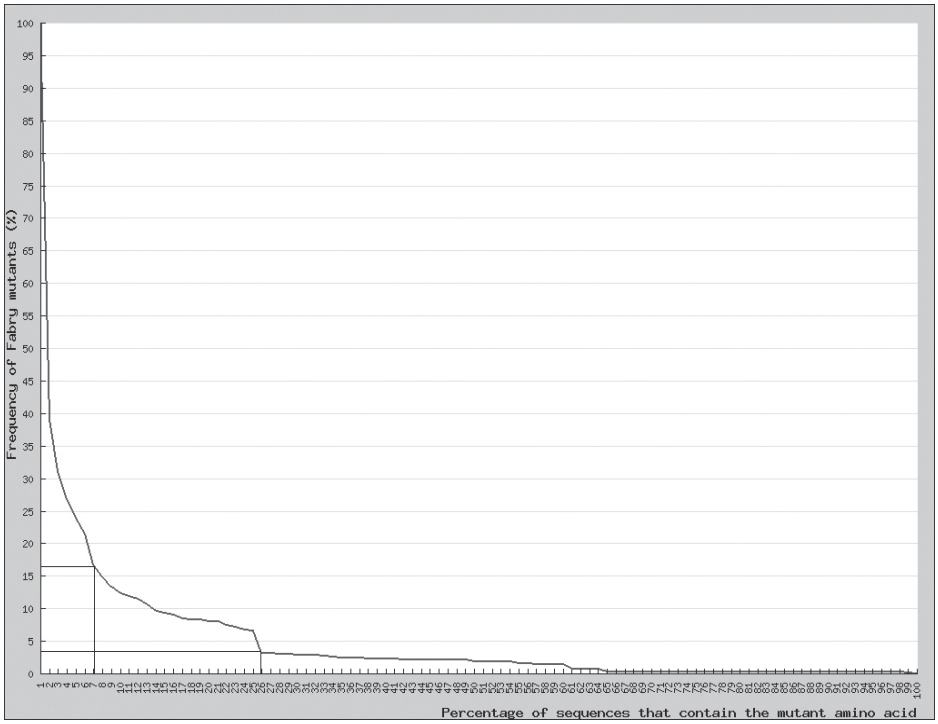


Figure 4. Correlation between the relative amino acid conservation (x-axis) and frequency of reported Fabry related mutations (y-axis). The x-axis represents the percentage of sequences that contains the mutated residue. The y-axis represents the percentage of the total number of fabry related mutation collected by Mutator. This plot shows that Fabry disease is most often the result of a mutation in *GLA* that resulted in a residue that is not commonly observed at the corresponding alignment position. Obviously there is a clear relation between the frequency of reported Fabry related mutations and the occurrence of amino acids at alignment positions.

3.3 Mutation analysis

1 Fabry disease-causing mutations and amino acid occurrences.

Super-family alignments can be considered as inventories of nature's successful mutagenesis experiments conducted during millions of years of evolution. In theory, the spectrum of residues present at a specific alignment position could be considered as allowed substitutions. Statistical analysis of super-family alignments can therefore potentially be used to predict the pathogenicity of specific mutations. This idea was tested using the set of Fabry related mutations collected in the FMDB. Figure 4 shows the relationship between the relative occurrence of amino acids at core positions and reported corresponding *GLA* mutations. For example, only 4% of the 1,117 reported *GLA* mutations in the core are mutations to an amino acid residue that is present at the corresponding alignment position in more than 26% of the aligned alpha-amylase sequences. Conversely, only 17% of mutations reported in structural conserved residues are mutated into an amino acid present at the corresponding alignment position in more than 7% of the aligned alpha-amylase sequences. Thus, the introduction of a new residue type that is infrequently observed in the complete alignment of the super-family at the particular alignment position has a high probability to be pathogenic implicating that this correlation can in principle be used to predict the pathogenicity of an unclassified variant (UV) in *GLA*. For example, if a particular UV is a mutation to an amino acid that is present in more than 25% of the alpha-amylase sequences at the corresponding alignment position, then the analysis suggests a small probability for pathogenicity for this particular UV. On the other hand, when the particular UV is present in less than 5% of the alpha-amylase sequences at the corresponding alignment position, then the analysis suggests a high probability for pathogenicity for the particular UV.

This correlation is not valid for non-core positions. For these positions only the amino acid occurrences of the 77 sequentially related sequences of the of *GLA* subfamily can be used. However, even within this small set, mutations at highly conserved positions are more likely to be pathogenic (see examples below).

2 Fabry disease-causing mutations and solvent accessibility

Solvent accessibility is the degree to which a residue in a structure is solvent exposed (e.g. more at the surface of the structure). Using a limited dataset of 278 missense mutations Garman [26] has shown that there is a strong correlation between solvent accessibility of residues and observed Fabry disease-causing mutations. The substantially increased mutational data collected in this study and the availability of the structural alignment makes it possible to study the predictability of solvent accessibility both at structurally conserved and non-conserved positions (Fig 5). Two correlations are plotted: 1) The correlation between Fabry disease-causing mutations at structurally conserved core positions and their solvent accessibility and 2) correlation between Fabry disease-causing mutations at structurally non-conserved positions and

their solvent accessibility. The plot clearly shows that this strong correlation exists specifically, almost exclusively, for core positions. This surprising observation is very important because it suggests that solvent accessibility should be used as indicator for pathogenicity only at core positions.

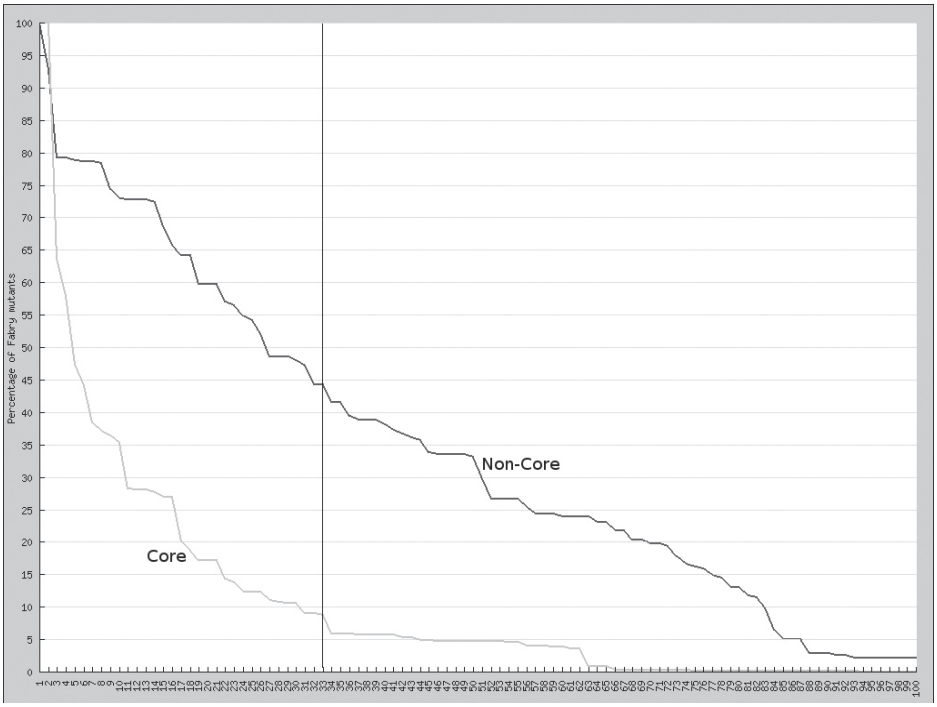


Figure 5. The correlation between Fabry disease-causing mutations at structurally conserved and non-conserved core positions and their solvent accessibility. X-axis: Percentage of accessible side chain surface area for each residue in the human *GLA* protein. Y-axis: Percentage of the total set of Fabry disease-causing mutations. Two plots are drawn. 1) The correlation between Fabry disease-causing mutations at structurally conserved core positions and their solvent accessibility (light grey line) and 2) correlation between Fabry

disease-causing mutations at structurally non-conserved positions and their solvent accessibility (dark grey line). The vertical black line indicates that only 7% of the 1,117 Fabry mutations located in the core are at positions of which the solvent accessibility > 33%. In contrast, almost half (44%) of the Fabry mutations located outside the core are at positions of which the residue has a solvent accessibility of > 33%.

3 Structural analysis of a specific amino acid change.

Validator performs a conformational analysis based on an estimation of the steric hindrance between the mutated residue and neighbouring residues in the 3D-structure. For that it generates an *in silico* model of the protein highlighting the substituted position including its van der Waals surface.

These types of analysis are done by Validator for each mutation uploaded to the FMDB website. The outcomes of other in DNA diagnostics commonly used classifiers (e.g. Grantham scores, BLOSUM62 scores) are also reported. Furthermore, amino acid specific information is provided, such as domain interface residue, active site residue and substrate contact information. Combining these predictions can lead to a better prediction.

For example, Fabry disease associated publications often report for 3D core position 76 (Ala143 in the *GLA* primary sequence) p.A143P. This mutation predisposes to a classical phenotype in males[27]. A statistical analysis of the complete super-family alignment indicates that at this position proline is the most abundant amino acid residue being present in 43% of the alpha-amylase sequences. The relative high solvent accessibility of Ala143 in *GLA* also suggests a low probability for pathogenicity at this site. Despite the above, the *in silico* model structure however, clearly indicates that specifically in *GLA* a proline at this position clashes with the neighboring aspartic acid with 3D-number 32 (Asp93 in *GLA*) (Fig. 6). Therefore, it is more likely that the p.A143P substitution is not allowed in the *GLA* protein and therefore can be considered as probably pathogenic. In contrast, Validator suggests that p.A143T would structurally be less damaging and has been reported to lead to a much milder variant of Fabry disease[27]. In this case the statistical analysis of the structural super-family alignment suggests pathogenicity since a threonine is seldom present (0.8%) in other alpha-amylase protein super-family members. The fact that p.A143T would structurally be less damaging fits well with a milder phenotype.

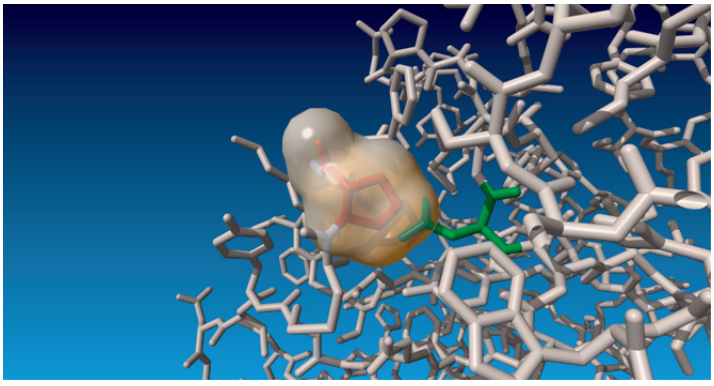


Figure 6. YASARA ball and stick backbone visualization of mutation p.A143P in GLA. The alanine to proline substitution at 3D-position 76 is depicted in red and surrounded by its Van der Waals surface. Clearly steric hindrance is observed with the side-chain of aspartate 93.

4 Performance of Validator tool on classical Fabry mutations.

To test the performance of Validator predictions, mutations known to result in the classical form of Fabry were selected being mutations, p.M42V, p.H46Y, p.D92Y, p.R112C, p.C142R, p.W226R, p.N320Y at core positions and p.P40S p.R100T at non-core positions. In addition, the special case p.D313Y is discussed.

For p.H46Y, Validator predicts a high probability for pathogenicity. In the super-family alignment, the occurrence of Y is only 4.1% and Figure 5 shows that more than 75% of the recorded Fabry mutations are the result of such a substitution. Also, the solvent accessibility is 1.5% indicating that the H46 is buried inside the protein. Furthermore, the *in silico* model suggests that p.H46Y causes bumps with surrounding amino acids. Since histidine residues are hydrophilic, a buried histidine almost always has an important function. Although currently no weight is given to the various indicators for this buried histidine solvent accessibility is probably the most important indicator.

Arguments listed above for p.H46Y are also true for p.D92Y. The fact that both position H46 and D92 are reported in more than 10 independent Fabry disease associated publications reporting substitutions to a number of different amino acids clearly match Validator predictions.

R112 is an almost completely buried hydrophilic residue. Almost all other sequences of the super-family have hydrophobic residues at this position instead. This indicates that R112 has an important function which is specific for *GLA*, which suggests that p.R112C will most probably be pathogenic. Furthermore, a cysteine is not a common residue at position 112 (0.1%) and the high number of publications (81) that report this position in relation to Fabry's disease again indicate a very high probability for pathogenicity.

The Validator tool indicates that C142 forms a cysteine bridge. The p.C142R mutation therefore disrupts the formation of this cysteine bridge. This type of information will overrule all others, since disrupting a cysteine bridge will most probably always be pathogenic independent of solvent accessibility, amino acid occurrences or other factors. Finally, almost all information that the Validator tool returns for mutations p.W226R and p.N320Y indicate a very high probability for pathogenicity again supported by a high number of publication reporting mutations to various amino acid types.

Mutations p.P40S and p.R100T are not included in the core, so only the 77 sequences of the *GLA* sub-family alignment can be used for statistics. In the sub-family both P40 and R100 are 100% conserved which suggests a high probability for pathogenicity for both.

The only mutation that is predicted not to be pathogenic is p.M42V. Even after meticulous manual inspection of the protein model of p.M42V, no reasonable explanation can be given for the pathogenicity of this mutation. The only indication that this is a true pathogenic mutation is the high number of literature references that report mutations to different amino acids at this position.

In the literature mutation p.D313Y is ambiguously linked with Fabry disease and the prediction from the 3DM data is contradicting. Although tyrosine is not a common residue at this position (suggesting a high chance for pathogenicity) solvent accessibility

indicates that the residue is located on the outside of the protein and introducing a tyrosine residue does not cause any bumps with surrounding amino acids (suggesting low probability for pathogenicity). The p.D313Y mutation has been tested for activity *in vitro*. Transient expression of the p.D313Y construct in COS-7 cells resulted in an active enzyme with >67% of the expressed wild type activity[28]. Mutator extracted 17 different publications from the literature all describing the single p.D313Y mutation but remarkably so far no other substitutions have been detected. Could this then be a naturally occurring variant? There are 46 other residues in the *GLA* protein sequence for which more than 10 independent Fabry disease related literature references are available. These are for residues 34, 40, 42, 46, 49, 22, 65, 66, 89, 92, 93, 97, 100, 112, 113, 138, 142, 143, 148, 156, 162, 172, 183, 205, 215, 220, 223, 226, 227, 236, 259, 266, 272, 279, 287, 296, 298, 301, 317, 320, 328, 342, 356, 357, 358, and 409. In contrast with reports for position D313 for all these positions, except for R220, a range of amino acid changes are reported. For R220 all 21 available independent publications report a stopcodon at position 220 (p.R220X). The fact that at these 46 positions different amino acid substitutions have been reported to result in Fabry disease significantly increases the chance that mutations at these positions are pathogenic. Furthermore, this result also indicates that p.D313Y is probably a naturally occurring variant, since it is unlikely that only the introduction of a tyrosine results in Fabry disease.

The results for the A143T and D313Y mutations fit whit what is clinically observed. Authors who report D313Y should comment that it is unlikely (but possible) to be pathogenic.

In this paper it is shown that a collection of super-family data can be used to predict effects of mutations. It must be noted, however, that predicting the pathogenicity of specific mutations is still difficult and statistical analysis of large 3DM alignments should only be used as guidance. For example, if we take the seven core positions that are conserved in more than 95% of the aligned sequences (3d-numbers 39, 73, 100, 102, 123, 145, 146) we see that for two of these positions (73, 123) Mutator has not been able to extract from the literature any mutations causing Fabry disease. Is this unexpected result caused by a still limited set of mutations or do mutations at these positions not lead to Fabry disease?

AUTHOR CONTRIBUTIONS

HJ and PS conceived the project. TB designed and developed the Mutator software with suggestions from HJ. RK performed the data analysis with suggestions from HJ and TB. RK and TB wrote the chapter with input from HJ. HJ, RH, MM and PS supervised the project.

REFERENCES

1. Petitjean, A., et al., *Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: Lessons from recent developments in the IARC TP53 database.* Human Mutation, 2007. **28**(6): p. 622-629.
2. Olivier, M., et al., *The IARC TP53 database: New online mutation analysis and recommendations to users.* Human Mutation, 2002. **19**(6): p. 607-614.
3. Horn, F., L. Lee, and F. Cohen, *MuteXt: An automated method to extract mutation data from the literature.* Pacific Symposium on Biocomputing, 2003.
4. Durme, J., et al., *NRMD: Nuclear Receptor Mutation Database.* Nucleic Acid Research, 2003. **31**(1): p. 331-333.
5. Horn, F., A. Lau, and F. Cohen, *Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors.* Bioinformatics, 2004. **20**(4): p. 557-568.
6. Stenson, P., et al., *The Human Gene Mutation Database: 2008 update.* Genome Medicine, 2009. **1**(1): p. 13.
7. Yip, Y.L., et al., *Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase.* Hum Mutat, 2008. **29**(3): p. 361-6.
8. Folkertsma, S., et al., *A Family-base approach reveals the function of residues in the Nuclear Receptor ligand-binding domain.* Journal of Molecular Biology, 2004. **341**(2): p. 321-335.
9. Joosten, H.-J., et al., *Identification of fungal oxaloacetate hydrolase within the isocitrate lyase/PEP mutase enzyme superfamily using a sequence marker-base method.* Proteins, 2008. **70**(1): p. 157-166.
10. Kuipers, R., et al., *Correlated mutation analyses on super-family alignments reveal functionally important residues.* Proteins, 2009. **76**(3): p. 608-616.
11. Leferink, N., et al., *Identification of a gatekeeper residue that prevents dehydrogenases from acting as oxidases.* Journal of Biological Chemistry, 2009. **284**(7): p. 4392-4397.
12. Narayan, B., et al., *Structure and function of 2,3-Dimethylmalate Lyase, a PEP Mutase/Isocitrate Lyase Superfamily member.* Journal of Molecular Biology, 2009. **386**(2): p. 486-503.
13. Folkertsma, S., et al., *The Nuclear Receptor Ligand-Binding Domain: A family-based structure analysis.* Current Medicinal Chemistry, 2005. **12**(9): p. 1001-1016.
14. Kuipers, R., et al., *3DM: systematic analysis of heterogeneous super-family data to discover protein functionalities.* Proteins, 2010. **Accepted**.
15. Murzin, A., et al., *SCOP: A structural classification of proteins database for the investigation of sequences and structures.* Journal of Molecular Biology, 1995. **247**(4): p. 536-540.
16. Altschul SF, G.W., Miller W, Myers EW, Lipman DJ, *Basic local alignment search tool.* Journal of Molecular Biology, 1990. **214**: p. 403-410.
17. Berman HM, W.J., Feng Z, Gilliland G, Bhat TN, Weissig H, Shindayalov IN, Bourne PE, *The Protein Data Bank*. Nucleic Acid Research, 2000. **28**(1): p. 235-242.
18. Boeckmann B, B.A., Apweiler R, Blatter M-C, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M, *The Swiss-Prot protein knowledgebase and its supplement TrEMBL*. Nucleic Acid Research, 2003. **31**(1): p. 365-370.
19. Garman, S. and D. Garboczi, *The molecular defect leading to Fabrys disease: Structure of human alpha-galactosidase.* Journal of Molecular Biology, 2004. **337**(2): p. 319-335.

20. Oliveira L, Paiva ACM, and V. G, *Correlated mutation analyses on very large sequence families.* chembiochem, 2002 **3**(10): p. 1010-7.

21. Grantham, R., *Amino acid difference formula to help explain protein evolution.* Science, 1974. **185**(4154): p. 862-4.

22. Henikoff, S. and J.G. Henikoff, *Amino acid substitution matrices from protein blocks.* Proc Natl Acad Sci U S A, 1992. **89**(22): p. 10915-9.

23. Sunyaev, S., V. Ramensky, and P. Bork, *Towards a structural basis of human non-synonymous single nucleotide polymorphisms.* Trends Genet, 2000. **16**(5): p. 198-200.

24. Ng, P.C. and S. Henikoff, *SIFT: Predicting amino acid changes that affect protein function.* Nucleic Acids Res, 2003. **31**(13): p. 3812-4.

25. Chafa, O., et al., *Characterization of a homozygous Gly11Val mutation in the Gla domain of coagulation factor X.* Thrombosis Research, 2009. **124**(1): p. 144-148.

26. Garman, S.C., *Structure-function relationships in alpha-galactosidase A.* Acta Paediatr Suppl, 2007. **96**(455): p. 6-16.

27. Benjamin, E.R., et al., *The pharmacological chaperone 1-deoxygalactonojirimycin increases alpha-galactosidase A levels in Fabry patient cell lines.* J Inherit Metab Dis, 2009. **32**(3): p. 424-40.

28. Froissart, R., et al., *Fabry disease: D313Y is an alpha-galactosidase A sequence variant that causes pseudodeficient activity in plasma.* Mol Genet Metab, 2003. **80**(3): p. 307-14.

CHAPTER 3

Common pitfalls and novel opportunities for predicting variant pathogenicity

Tom van den Bergh, Bas Vroling, Remko KP Kuipers, Henk-Jan Joosten and Gert Vriend

BIOCHEMISTRY & PHYSIOLOGY: OPEN ACCESS 2016, FEB;5:197

ABSTRACT

The prediction of missense variant pathogenicity is normally performed using analyses of multiple sequence alignments, optionally augmented with analyses of the (predicted) protein structure. The most straightforward way, though, is to search the literature to see whether this variant has already been described. Variant data from homologous proteins are also valuable because mutations in a homologous protein often have similar effects as mutations at the equivalent residues of the protein of interest. Transferring variant data seems trivial but is seriously hampered by the fact that homologous residue positions have different numbers in different species. This problem is even bigger when to proteins have such low sequence identities that they can no longer be aligned based on their sequences only and their structures need to be compared to align them accurately. The protein superfamily analysis software suite 3DM solves these problems, because 3DM is a system that combines high quality structure based multiple sequence alignments in which aligned residues have the same number, with all published mutant and variant data for human and all other species. We have used 3DM to analyze nine human proteins for which many disease-related variants are known. This study reveals that mutation data can be transferred even between very distant homologous proteins. Thus, protein superfamily information systems, such as 3DM, offer a wealth of unused information that can be used in the analysis of human variants.

1 INTRODUCTION

Rapidly evolving gene sequencing technologies have revealed the relation between mutations in genes and the onset of symptoms that can be assigned to corresponding genetic disorders. More than a hundred thousand unique pathogenic variants are available from the Human Gene Mutation Database (HGMD) [1] that are known to cause over ten thousand different monogenic disorders [2] and almost four thousand genes already have been described that are involved in polygenic disorders [3]. Next generation gene sequencing efforts, on the other hand, have revealed that harmless single nucleotide polymorphisms (SNPs) are even more frequent in the human genome. More than ten million DNA variations have been uncovered in the human genome, of which about 4% are located in gene coding regions and about half of those (2%) are non-synonymous SNPs (nsSNPs) that thus result in an amino acid change in the corresponding proteins [4]. In fact, it was estimated by Crawford et al. that on average each gene contains five nsSNPs that are present in more than 5% of the human population. Only a small fraction of these nsSNPs have an effect on the function of the corresponding protein and can be classified as pathogenic. It is evident that gene sequencing is a promising method for diagnosis of genetic disorders, but the frequent occurrence of benign variants drastically hampers routine diagnosis of genetic disorders.

Mutation databases, such as HGMD, OMIM [5], and protein specific databases, such as for example the P53 mutation databases [6,7], can be used as reference for previously identified pathogenic variants. In the daily practice of DNA diagnostics, one obviously encounters many variants for which these databases have no information available yet. In such cases one normally resorts to literature searches or, if that fails, to any of a series of tools, such as Polyphen-2 [8], SIFT [9], and HOPE [10] that have been developed for the *in silico* evaluation or prediction of the effects of variants on gene and protein function. However, in many cases variant effects are known for homologous proteins. The concept of evolution is that all homologs share a common ancestor and thus it makes sense that variants on equivalent positions would cause a similar effect (e.g. cause a disease).

The first step in transferring mutability data between proteins is to identify the equivalent positions. However, to confidently determine which positions are equivalent is difficult unless the proteins in the alignment are so similar that aligning them becomes trivial. Jordan *et al*. [12] showed that the publicly available mutant severity prediction methods sometimes produce very poor alignments, and thus questionable predictions. Therefore, protein superfamily information systems that use structural alignments are needed to analyze these proteins and enable data transfer between proteins that are more distantly related but still share a common fold.

We have previously described 3DM[13], a system that can generate superfamily systems that fully integrate sequence data with literature data, mutation information, and three-dimensional structures. The 3DM multiple sequence alignments are derived from structure superpositions. This makes them more correct than commonly available alignments, and it allows for larger numbers of sequences to be reliably included in the alignments. 3DM systems contain all available sequence variants (protein- and DNA sequences of all splice variants, with and without leader sequences). We have generated 3DM systems for five proteins that are involved in long QT syndrome (gene names: KCNQ1, KCNH2, SCN5A, SCN1A, KCNJ2) and for four members of the amylase superfamily that are involved in Fabry disease, Schindler- or Kanzaki disease, glycogen storage disease, and cystinuria (gene names: GLA, NAGA, GBE1, SLC3A1, respectively).

In this work we show that the mutability of residues can still be transferred even on a superfamily scale where proteins are sequentially very different. The sequence identity between the proteins in our manuscript is as low as 10% and thus these proteins can no longer be aligned by sequence alignment programs. Therefore, protein superfamily information systems that use structural alignments, such as 3DM, are needed to analyze these proteins and transfer data for pathogenicity predictions. Additionally, we show that automated literature mining software can outperform manually curated databases such as HMGD, both in terms of the number of unique mutations extracted, as well as the depth of information per mutation.

2 METHODS

2.1 3DM information systems

The 3DM software that generates superfamily information systems is extensively described elsewhere [13,14], and will here only be discussed briefly. A structure based multiple sequence alignment (MSA) forms the backbone of each information system. 3DM uses protein structure data to determine which regions are structurally conserved. These regions are termed core regions and 3DM normally uses only these superfamily core regions to generate the superfamily alignments. All sequences and structures are renumbered so that residues aligned in the MSA get the same number throughout the information system. This enables the transfer of data and knowledge between proteins and facilitates literature searches for mutations in homologs.

2.2 Multiple sequence alignments

To predict the pathogenicity of non-synonymous variants the quality of alignments is of much greater importance than the completeness of the MSA. Structure based sequence alignment methods, which are used by default in the 3DM systems, tend to produce alignments of higher quality and deeper coverage than classical MSA methods. However, for the long QT related genes, structure information is available for only small parts of the proteins. To enable high-quality predictions, we implemented a method

that aligns only those parts of the sequences that can be aligned with great confidence, similar to the way the PROTOMAT algorithm produces what they call BLOCKS: ungapped regions of aligned proteins [15]. For the parts of the long QT related proteins for which structural information is available, structure-based alignments were produced and were subsequently merged with the sequence-based MSAs.

2.3 Mutation data

Mutations were extracted from the literature by the 3DM Mutator module [14]. PubMed was queried for papers containing mutations related to the protein members of the two protein superfamilies here investigated. For the alpha-amylase superfamily Mutator scanned 11,471 full-text papers, whereas mutation-related information for the potassium channel superfamily was extracted from 41,253 full-text articles. In total, this resulted in 5,219 and 65,891 mutations for the alpha-amylase and potassium channel superfamilies, respectively.

3 RESULTS

We investigated the transferability of several types of information among members of protein superfamilies, and the power and limitations of automatically extracting mutation data from the literature.

Table 1. The overlap of mutations between different disease related proteins.

Protein 1	Protein 2	Core identity	Aligned positions	Mutations protein 1	Mutations protein 2	Overlap	p-value
GLA	SLC3A1	0.15	210	150	37	31	0.046
GLA	NAGA	0.57	209	149	8	8	0.062
GLA	GBE1	0.10	194	140	12	8	0.787
GLA	SLC3A1, NAGA	-	210	150	45	39	0.006
GLA	SLC3A1, NAGA, GBE1	-	210	150	57	47	0.021
KCNQ1	KCNH2	0.17	156	109	126	96	0.001
SCN5A	SCN1A	0.73	1576	476	271	101	0.003

From left to right the columns show the two proteins of the comparison; the sequence identity of the aligned region; the number of aligned positions; the number of positions in both proteins at which pathogenic mutations have been observed, the number of those that overlap, and the p-value for the overlap to be random determined by a permutation analysis.

3.1 Pathogenic variants tend to cluster at equivalent positions

The basic assumption that allows for the transfer of mutation data between protein family members is that mutations at equivalent positions in homologous proteins tend to result in similar structural and functional effects. Based on this assumption it is to be expected that structurally related proteins have equivalent locations where mutations are well tolerated and locations where mutations are prone to result in detrimental effects. We have tested this hypothesis by analyzing the pathogenic variants, that were extracted from the literature by Mutator, in the two protein sets. None of the pathogenic variants are observed with a minor allele frequency of 1% or higher in the ExAC population database [16,17]. The first test set consists of 381 aligned pathogenic variants in four human proteins of the α-amylase superfamily (GLA, NAGA, GBE1, and SLC3A1) that, when mutated, can result in Fabry disease, Schindler- or Kanzaki disease, glycogen storage disease, or cystinuria, respectively. These proteins are sequentially very distantly related to each other. Sequence identities of these proteins range between 10% and 57% as shown in Table 1, which makes it (almost) impossible to align them correctly using sequence based alignment method that are normally used by the standard variant prediction tools (e.g. SIFT or PolyPhen [8,9]), but due to their similar protein structures they can still be aligned correctly (see figure 1). The 381 pathogenic variants are observed at 158 structurally different positions. Figure 2 shows how often pathogenic variants are observed at corresponding positions in these four proteins.

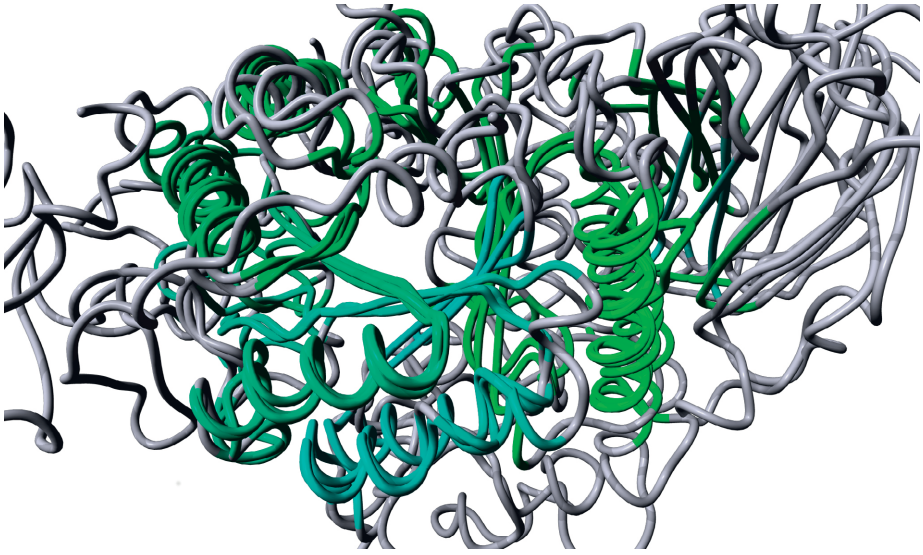


Figure 1. Structural alignment of subfamily representative structures for the four human proteins in the alpha amylase superfamily. GLA and NAGA are both represented by 1R47 chain B, SLC3A1 is represented by structure 2DH3 chain A, and GBE1 is represented by structure 1M7X chain A. The blue to green regions are considered structurally conserved while the gray regions are structurally variable in this alignment. The blue to green gradient visualizes the order of the conserved regions from the N- to C-terminus respectively.

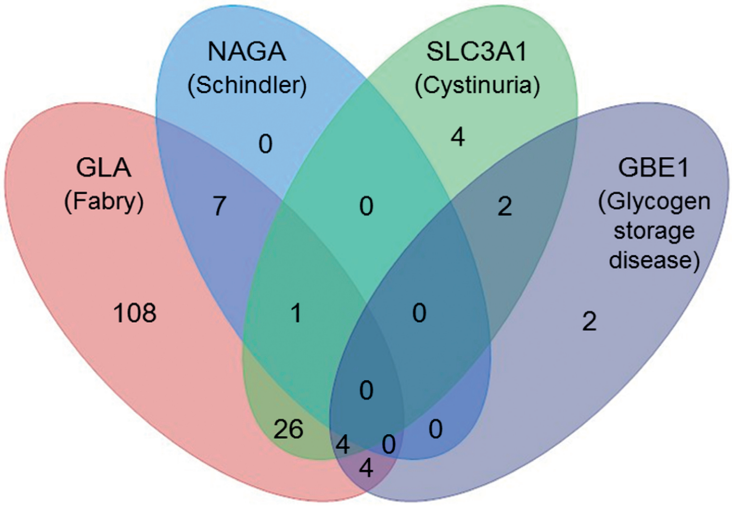


Figure 2. The relation of pathogenic mutations of different homologous proteins and their corresponding diseases. Overlapping parts of the ovals represent equivalent protein positions and the numbers are the number of positions for which mutation data is detected. For instance, there are 31 positions for which pathogenic mutations have been detected in both GLA and SLC3A1 proteins.

For our second test set, the four potassium channels, structural information is present only for a very small fraction of these proteins, which hampers the transfer of mutation data between these proteins. SCN5A and SCN1A are sequentially closely related and these two can reliably be aligned over nearly the full lengths of their sequences. However, KCNQ1 and KCNH2 can only reliably be aligned at 156 positions that are structurally conserved. This is the transmembrane region of these proteins. No structure data is available for SCN5A and SCN1A and these proteins can reliably be aligned to KCNQ1 and KCNH2 at only 29 of these 156 positions. Due to the absence of structural information and the limited number of mutation data for the four potassium channels, the significance of the overlap of mutation data could only be determined for SCN5A and SCN1A and for the 156 structural conserved positions of KCNQ1 and KCNH2. For these two datasets an even greater overlap is observed of positions that are disease related in both families. Table 1 provides the numerical details of these analyses.

3.2 Variation in close homologs is indicative of mutation-tolerant positions

It is commonly accepted that mutations at conserved positions are likely to be pathogenic, and many MSA-based software packages (e.g. SIFT; [9]) that aim at predicting the significance of mutations for a disease state implicitly use this concept. If an alignment consists only of highly similar sequences, then obviously most positions will be observed as conserved. If in a sequence alignment all sequences are more than 90% sequence identical to the human sequence then obviously the MSA contains only se-

quences from species that are closely related to homo sapiens, and consequently, any variability observed in this MSA is likely to also be acceptable in the human sequence. To test this hypothesis, we compared the ratio of pathogenic variants at conserved positions with the ratio of pathogenic variants at non-conserved positions. To ensure that this test was statistically meaningful we only used two of the nine human proteins (one from each super-family) for which a large number of different pathogenic variants (>250) are available. Table 2 shows that pathogenic variants are less frequently observed at variable positions in alignments composed of only highly similar sequences. For instance, for 179 of the 420 positions in the alignment of closely related GLAs pathogenic variants (stop codons and deletions excluded) have been reported in the HGMD database. The alignment composed of sequences that are at least 90% identical to GLA contains 387 conserved positions. For 176 (45%) of these positions pathogenic mutations have been described. For the variable positions this is five times less, because for only 3 of the remaining 33 variable positions (9%) pathogenic variants have been described. To determine the significance of the lower frequency of pathogenic mutations at variable positions a p-value was determined, which was <0.01 for all factor values from table 2.

Table 2. The relation between pathogenicity of mutations at conserved positions versus variable positions.

	GLA		KCNQ1	
Identity	**factor**	**p-value**	**factor**	**p-value**
0.9	5.0	< 0.001	2.24	0.0045
0.85	4.0	< 0.001	2.25	< 0.001
0.8	3.5	< 0.001	2.07	< 0.001
0.75	3.2	< 0.001	1.98	< 0.001
0.7	2.8	< 0.001	2.07	< 0.001
0.65	2.8	< 0.001	2.00	< 0.001
0.6	2.6	< 0.001	1.91	< 0.001

The left column represents the sequence identity compared to the human sequence. The factor column indicates how much more often pathogenic mutations are found at 100% conserved positions than at variable positions. For instance, using an alignment composed of sequences that are 90% or more identical to GLA this factor is 5.0, which means that the percentage of conserved positions at which pathogenic mutations have been observed is 5.0 times higher than positions at which at least one of the aligned homologs has a different residue type than the human sequence. Clearly, human mutations are more easily tolerated at positions that are variable in highly related sequences.

3.3 Mutation data extraction: automated mining outperforms manually curated approaches

HGMD [1] is the *de facto* standard source for mutation information. Like any manually curated database, the high quality of the HGMD comes at the cost of incompleteness. It was shown recently that Mutator is able to extend the HGMD [18]. Figure 3 shows that Mutator extracts significantly more mutations from the literature than human experts. This does, of course, not invalidate systems like HGMD because the HGMD also provides details on the effect of a variant. Even though natural language parsing software is developing very rapidly the days that the successor of Mutator will also always correctly extract the effects of those mutations from the literature are not near. In contrast to the HGMD database that stores each unique mutation once, 3DM collects all publications that describe any particular mutation. This can, for example, be two publications that describe the same mutation detected in different patients with different onsets or different symptoms. We have frequently observed that contradicting information is reported for the same mutation in different patients.

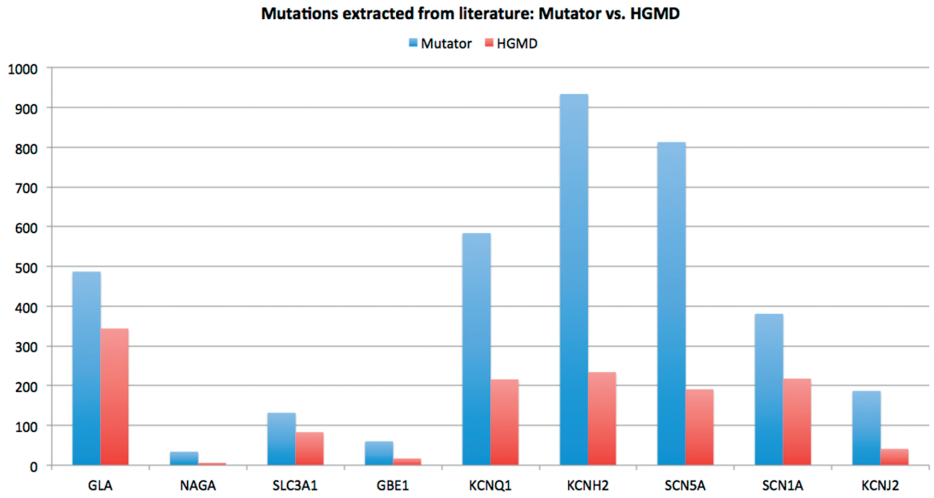


Figure 3. Comparison of the number of unique mutations extracted by Mutator and HGMD for the α-amylase and LQT protein family members.

4 DISCUSSION

We have made a number of interesting observations. First, we find that variants are more likely to be pathogenic if they occur in structurally conserved regions of a protein [14]. Second, we find that it is much less likely that variants are pathogenic if they occur at positions that are variable in alignments composed of only highly similar proteins. From this observation follows that if a close homolog of a human protein has a different residue, it is more likely that other residue types are allowed at the equivalent position of the human protein. Third, we show that there is a large overlap in alignment positions

where pathogenic mutations occur even among distantly related human proteins of a superfamily. Therefore, when a missense mutation is pathogenic to its host organism, the chance that a mutation at the equivalent position in a homologous protein (to any residue) is also pathogenic is much higher. These observations show, as was hinted at previously [10], that pathogenicity is much more determined by the location of the mutation in the protein than by the type of amino acid that is introduced. We can conclude that the use of protein superfamily systems can extensively add previously unused data for the investigation of human disease related variants. These revelations can function as very useful predictive features for variant effect prediction models. The availability of a protein superfamily data integration system is valuable for such a model, since it can provide predictive features that otherwise would be missing, such as mutation data for very distant homologs. In fact, these models have been generated for the LQT related genes. We show that the use of superfamily data, largely increases the accuracy of variant effect predictions (publication in progress).

AUTHOR CONTRIBUTIONS

Conceptualization: TB HJ. Data curation: TB RK. Formal analysis: TB. Funding acquisition: HJ. Investigation: TB. Methodology: TB BV. Project administration: HJ. Resources: HJ. Supervision: HJ GV. Validation: TB BV. Visualization: TB. Writing – original draft: TB. Writing – review & editing: TB BV.

REFERENCES

[1] Stenson PD, Ball EV, Howells K, Phillips AD, Mort M, et al. (2009) The Human Gene Mutation Database: providing a comprehensive central mutation database for molecular diagnostics and personalized genomics. Human Genomics 4: 69–72.

[2] World Health Organization. (2011) WHO | Genes and human disease [Internet].

[3] Cooper DN, Chen J-M, Ball EV, Howells K, Mort M, et al. (2010) Genes, mutations, and human inherited disease at the dawn of the age of personalized genomics. Human Mutation 31: 631–55.

[4] Crawford DC, Akey DT and Nickerson DA. (2005) The patterns of natural variation in human genes. Annual Review of Genomics and Human Genetics 6: 287–312.

[5] Hamosh A, Scott AF, Amberger J, Valle D and McKusick VA. (2000) Online Mendelian Inheritance in Man (OMIM). Human Mutation 15: 57–61.

[6] Béroud C and Soussi T. (2003) The UMD-p53 database: new mutations and analysis tools. Human Mutation 21: 176-81.

[7] Olivier M, Eeles R, Hollstein M, Khan MA, Harris CC, et al. (2002) The IARC TP53 database: new online mutation analysis and recommendations to users. Human Mutation 19: 607–14.

[8] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, et al. (2010) A method and server for predicting damaging missense mutations. Nature Methods 7: 248–9.

[9] Ng PC and Henikoff S. (2003) SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Research 31: 3812–4.

[10] Venselaar H, Te Beek TAH, Kuipers RKP, Hekkelman ML and Vriend G. (2010) Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. BMC Bioinformatics 11: 548.

[11] Gong S, Worth CL, Cheng TMK and Blundell TL. (2011) Meet Me Halfway: When Genomics Meets Structural Bioinformatics. Journal of Cardiovascular Translational Research.

[12] Jordan DM, Kiezun A, Baxter SM, Agarwala V, Green RC, et al. (2011) Development and validation of a computational method for assessment of missense variants in hypertrophic cardiomyopathy. American Journal of Human Genetics 88: 183–92.

[13] Kuipers RK, Joosten H-J, van Berkel WJH, Leferink NGH, Rooijen E, et al. (2010) 3DM: systematic analysis of heterogeneous superfamily data to discover protein functionalities. Proteins 78: 2101–13.

[14] Kuipers R, van den Bergh T, Joosten H-J, Lekanne dit Deprez RH, Mannens MM, et al. (2010) Novel tools for extraction and validation of disease-related mutations applied to Fabry disease. Human Mutation 31: 1026–32.

[15] Henikoff S and Henikoff JG. (1991) Automated assembly of protein blocks for database searching. Nucleic Acids Research 19: 6565–72.

[16] Exome Aggregation Consortium, Lek M, Karczewski K, Minikel E, Samocha K, et al. (2015) Analysis of protein-coding genetic variation in 60,706 humans. bioRxiv 030338.

[17] Song W, Gardner SA, Hovhannisyan H, Natalizio A, Weymouth KS, et al. (2015) Exploring the landscape of pathogenic genetic variation in the ExAC population database: insights of relevance to variant classification. Genetics in Medicine: Official Journal of the American College of Medical Genetics.

[18] Stenson PD and Cooper DN. (2010) Prospects for the automated extraction of mutation data from the scientific literature. Human Genomics 5: 1–4.

CHAPTER 4

CorNet: Assigning function to networks of co-evolving residues by automated literature mining

Tom van den Bergh[#], Giorgio Tamo[#], Alberto Nobili, Yifeng Tao, Tianwei Tan, Uwe T. Bornscheuer, Remko K. P. Kuipers, Bas Vroling, René M. de Jong, Kalyanasundaram Subramanian, Peter J. Schaap, Tom Desmet, Bernd Nidetzky, Gert Vriend, and Henk-Jan Joosten
 # These authors contributed equally

PLOS ONE 2017, MAY;12(5)

ABSTRACT

CorNet is a web-based tool for the analysis of co-evolving residue positions in protein super-family sequence alignments. CorNet projects external information such as mutation data extracted from literature on interactively displayed groups of co-evolving residue positions to shed light on the functions associated with these groups and the residues in them. We used CorNet to analyse six enzyme super-families and found that groups of strongly co-evolving residues tend to consist of residues involved in a same function such as activity, specificity, co-factor binding, or enantioselectivity. This finding allows to assign a function to residues for which no data is available yet in the literature. A mutant library was designed to mutate residues observed in a group of co-evolving residues predicted to be involved in enantioselectivity, but for which no literature data is available yet. The resulting set of mutations indeed showed many instances of increased enantioselectivity.

1 INTRODUCTION

The enormous progress in sequencing technology has increased the number of available sequences to hundreds of millions. For instance, the metagenome sequencing of just the biological diversity found in the Sargasso sea alone as reported by Craig Venter and coworkers[1] identified 1.2 million new genes. Within the *Global Ocean Survey* (GOS) project another 6.1 million new gene sequences were found. As shown by Rusch et al. (2007)[2,3] 1,700 new protein families could be discovered in these databases. This rich source of information are a gold mine for the life sciences as these genes encode for a plethora of novel and mostly unexplored enzymes useful for various areas such as medical science, pharmacy and biocatalysis[4].

During evolution, proteins undergo random mutations that leave their footprint in multiple sequence alignments (MSA). Some amino acid residues will stay conserved, others are conserved in groups of species, and yet others seem to mutate without restrictions. As a result we observe in multiple sequence alignments a hierarchy of residue conservation, correlation, and variation[5–7]. When residues are conserved within groups of sequences that share a certain function but these residues differ between groups we observe correlated mutation behaviour (also called co-evolution), and often such groups of residues are involved in a common function, such as specificity, co-factor binding, protein-protein interactions. We will call such groups of co-evolving residues 'networks'. CorNet is designed for the analysis of networks and for the prediction of their roles in protein function.

Many attempts have been made to use correlation patterns for the prediction of protein structures using information obtained from a MSA. Older methods all use what is now known as mutual information. A series of CASP[8] experiments illustrated that mutual information obtained from a MSA could not adequately predict protein structures. Recently a series of developments[9–12], have caused a breakthrough in the use of correlated mutations for the *ab initio* prediction of structures. Mutual information has often been related to function[13–16], and distinguishing correlated mutations reflecting residue contacts from those reflecting functions was the major problem faced when predicting protein structure from a MSA. These problems are not encountered, though, when studying or optimizing protein function in fields like protein engineering, chemical biology, or the analysis of disease causing mutations in the human exome because the strongest correlations, and especially whole networks of correlations often reflect a function[14].

Proteins have many functions including ligand and co-factor binding, regulation, signalling, membrane embedding and catalysis. Each function requires that a series of residues work together. Therefore, residues have not co-evolved in a pair-wise manner but rather as networks[17,18]. The concept of extracting correlated mutations from align-

ments is not new and many methods have been described previously[5,6,13–16,19,20]. Several correlated mutation analysis (CMA) software packages exist (e.g. ET[16], WHAT IF[21]) that cluster detected pairs of residues into networks. Networks are often composed of sub-networks each containing residue positions involved in one particular protein feature. A complicating factor in the assignment of residues to functions is that they often contribute to multiple functions [22,23].

The function of a network cannot be determined from physicochemical characteristics of the residues involved, but visual inspection of the 3D structure of the protein can reveal the function of a network. Fig 1 shows examples of networks in four super-families that surround ligand- and the cofactor binding pockets. Normally, though, the determination of function requires *in vitro* or *in vivo* experiments, but often such experiments have already been performed in either the molecule of interest or in a homolog and these results can often be extracted from the literature. Besides that the amount of available literature often is overwhelming, a literature study for the functional role of a residue can be complicated by the facts that residues often do not have the same numbers in close homologs and that proteins do not have the same names in different research fields. These problems have been solved in molecular class specific text-mining methods[24,25] that iterate between text analysis and validation using the MSA-based super-family information system.

Six protein super-family systems were used to demonstrate the relation between correlated mutation networks and mutation data that is available in the literature. These six super-families were chosen because they could be made available to the public. We show that very different functions can be the driving force behind the major network in a protein super-family. Specificity is the driving force in two of the six super-families, whereas we observe that twice co-factor binding, once activity, and once enantioselectivity lead to the strongest correlated mutations. Furthermore, we show that randomly deleting a large number of sequences from the input alignment hardly has an effect on the positions that make up the CorNet network. However, deleting entire groups of sequences that are phylogenetically closely related result in CorNet networks consisting of different alignment positions. In fact, when using alignments generated of carefully selected subsets of sequences the networks will reflect different functions compared to networks obtained from the whole super-family. We also show that the enrichment of residues involved in a certain function can be optimised by interactive modification of correlation cut-off values (enrichment is defined as the fraction of residues in the network that is related to the function relative to the fraction of residues related to that function in the whole protein). Enrichment factors between five and ten are not uncommon.

To validate if residues that are connected in a CorNet network indeed share a common function a targeted mutant library was designed for an α/β-hydrolases enzyme (the

Pseudomonas fluorescens esterase). We experimentally validated the relation between the major network and the associated keyword 'enantioselectivity'. The analysis of the residue distributions in this network allowed us to design a small library consisting of only 72 variants of which 18% showed a positive effect on enantioselectivity.

The explosion of readily available sequence- and mutation data is likely to make the type of protein data analysis described in this work a standard tool for scientific research in protein engineering and other protein related research fields.

2 RESULTS

With the CorNet server the user can select parameters such as correlation scores or colour schemes. The user can rapidly obtain information such as amino acid distributions at single positions or pairs of positions. In the six systems for which we performed the bibliome determination, the user can select search terms in that bibliome and results can be presented as scenes for visualisation of the CorNet data in a protein structure with the Yasara (www.yasara.org) macromolecular structure visualizer (Fig 1 shows examples).

CorNet is connected to the web based CMA tool Comulator and can be used by uploading an alignment to the embedded Comulator tool (www.bio-prodict.nl/comulator). CorNet is also part of the 3DM protein super-family analysis suite. For several publicly available 3DM systems, including the six 3DM databases described in this paper, the alignments, the CMA results, the CorNet networks (including the connection to Yasara), and the mutation data from the bibliome can be retrieved from www.3dm.bio-prodict.nl.

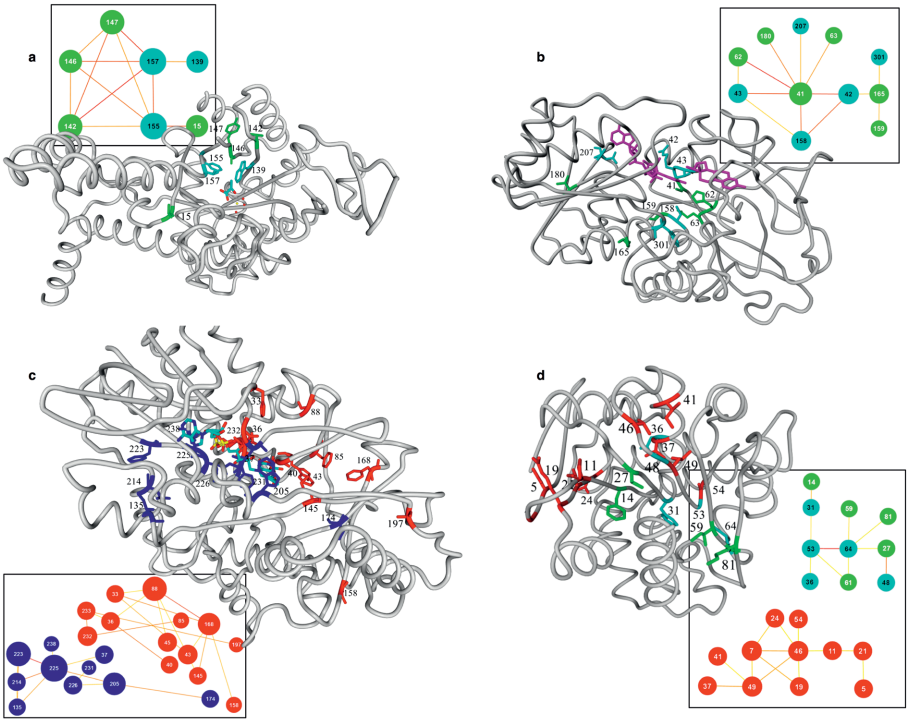


Fig 1. Visualisation of correlated mutation networks in the protein structures. In the boxes the correlated mutation networks are shown. Nodes represent alignment positions. Node sizes indicate the number of edges. Nodes shown in cyan indicate residue positions for which keyword related mutation data is available in the literature. Edge colours indicate the strength of the pair-wise correlation (yellow to red). The residues visualized in the structures correspond with, and match colours with nodes in the network.

a. Correlated mutation network of the isocitrate lyases (ICL) visualised in structure pdb-code: 1IGW. The cyan nodes in this network are related to the keyword 'specificity'. **b.** Correlated mutation network of the alcohol dehydrogenases (ADH) visualized in pdb-code: 1D1T that contains a substrate analog in the active site and the NAD co-factor (magenta). Position 41 that is the central hub in the correlation network is also the centre of the 3D network and it is located between the NAD and the substrate-binding pocket. **c.** Correlated mutation network of the amino acid oxidases (AAO) visualized in pdb-code: 1B37 that contains the FAD co-factor. This network consists of two sub-networks (blue surrounding the FAD, red surrounding the substrate binding pocket). **d.** Correlated mutation network of the a/b-hydrolase fold enzymes (a-bH) visualized in pdb-code: 1VA4. This network consists of two sub-networks. The smaller sub-network is highly enriched with positions (cyan nodes) related to the keyword 'enantioselectivity'.

CorNet was tested on six protein super-families: alcohol dehydrogenases (ADH), amino acid oxidase-like (AAO) proteins, RmlC-like cupin proteins (cupins), the phosphoenolpyruvate mutase/isocitrate lyases (ICL), UDP glycosyltransferases (UDP-GT), and a/b-hydrolases (a-bH). We used the Mutator tool[24] to extract from the literature the mutations associated with a series of functions including selectivity, activity, agonist binding, regulation, post-translational modification, and for validation purposes a series

of neutral terms such as stability, or the words 'the' and 'and'. In all six families we find that the strongest correlating network clearly relates to a main functional aspect.

2.1 Structural location of correlated mutation networks

Fig 1 shows the structural position of the correlated mutation networks of four super-families. Fig 1a-b shows the networks for the ICL and ADH superfamilies for which only a single significant network is observed. The AAO and a-bH families reveal a series of significant networks and Fig 1c-d show their locations in the respective 3D structures. Fig 1 allows for a series of observations. For example, there is a tendency for residue positions in the same network to also be located roughly in the same area in the 3D structure, but high CMA scores do not tend to relate to 3D contacts. In the AAO family all residues in the blue network are close to the FAD while most residues in the red network are in or near the active site. In none of the six networks do we see that residues that seem central (a hub) in the network are central in their 3D cluster too. The close spatial proximity of network residues seems caused simply by the fact that functions, such as catalysis or co-factor binding, are performed by residues that must lie around the active site or the co-factor. The conclusion that residues in a correlated mutation network will be involved in the same function is corroborated by experimental mutation studies for all of the six super-families studied here. These observations indicate that strongly correlated mutations in multiple sequence alignments are a result of functional constraints rather than structural contacts.

2.2 Residue function determination by enrichment

To find the function of residues in a CorNet network literature-extracted mutation data related to different keywords, such as 'specificity' and 'co-factor', were mapped on the network and the overrepresentation (enrichment) of these keyword-related mutations inside the network is determined by the calculation of enrichment score (Escores). The calculation of Escores is described in the Materials and Methods section. Fig 2 shows for each of these keywords this enrichment in relation to the correlation cut-offs. These enrichments are hard to quantify because of a series of reasons that range from bias in the main research topic in a certain field of the life sciences to low counting statistics caused by, for example, CMA networks reducing to just two amino acid positions at the highest CMA values. Another effect is that researchers tend to make mutations at 'positions of interest' and being interesting often is defined by literature describing mutations at that position in homologous proteins. We also observe large differences in the amount of mutation data available per super-family. Originally, we arbitrarily decided that mutations related to a selected keyword had to be observed in at least two independent articles before we would accept it as real. For the ICL and Cupin super-families, this 'two article' cut-off had to be abandoned to obtain any results. We do not have enough datasets available yet to start thinking about a relation between the number of available mutation articles, the length of the sequence, the number of sequences in the MSA, and the optimal cut-off for this parameter.

2.3 Enrichment scores

We measured the enrichment for a series of control keywords to at least get a qualitative idea about the significance of Escores. The control keywords 'and' and 'the' were selected because one expects these words to be observed frequently but randomly in sentences that are picked-up by the logical expressions that scan the literature for sentences that also contain the logical expression for a particular mutation (e.g. P213S). The Escores for these control keywords ranges between 0.00 and 2.02 in five of the super-families (Table 1). We also used the word 'stability' and 'zinc' as control keywords. Table 1 shows the enrichments for these four control keywords measured at a CMA value of 0.80. This value was chosen to ensure that the six super-families contain enough nodes to prevent biased enrichment scores, which can result from the fact that scientist tend to select 'interesting' positions to mutate. The Network of the ICL super-family is surrounding the active site. Therefore, this biased selection of amino acids results in enriched control keywords simply because there are only a limited number of experimental mutations available.

Table 1. Enrichment scores for control keywords[a].

keyword	and	the	stability	zinc
ADH	1.15	1.18	2.02	–
AAO	1.16	1.15	1.94	0.00
Cupin	1.00	0.94	0.87	1.40
ICL	4.04	3.83	0.00	1.00
UDP-GT	1.54	1.53	0.00	0.00
a-bH	0.00	0.00	1.35	1.57

[a] The enrichments were calculated at a CMA cut-off of 0.80. The keyword 'zinc' is not shown for the ADH super-family because zinc is a co-factor in this family and thus not a control keyword.

Fig 2 shows for the six super-families the relation between mutations and a series of keywords and their Escores.

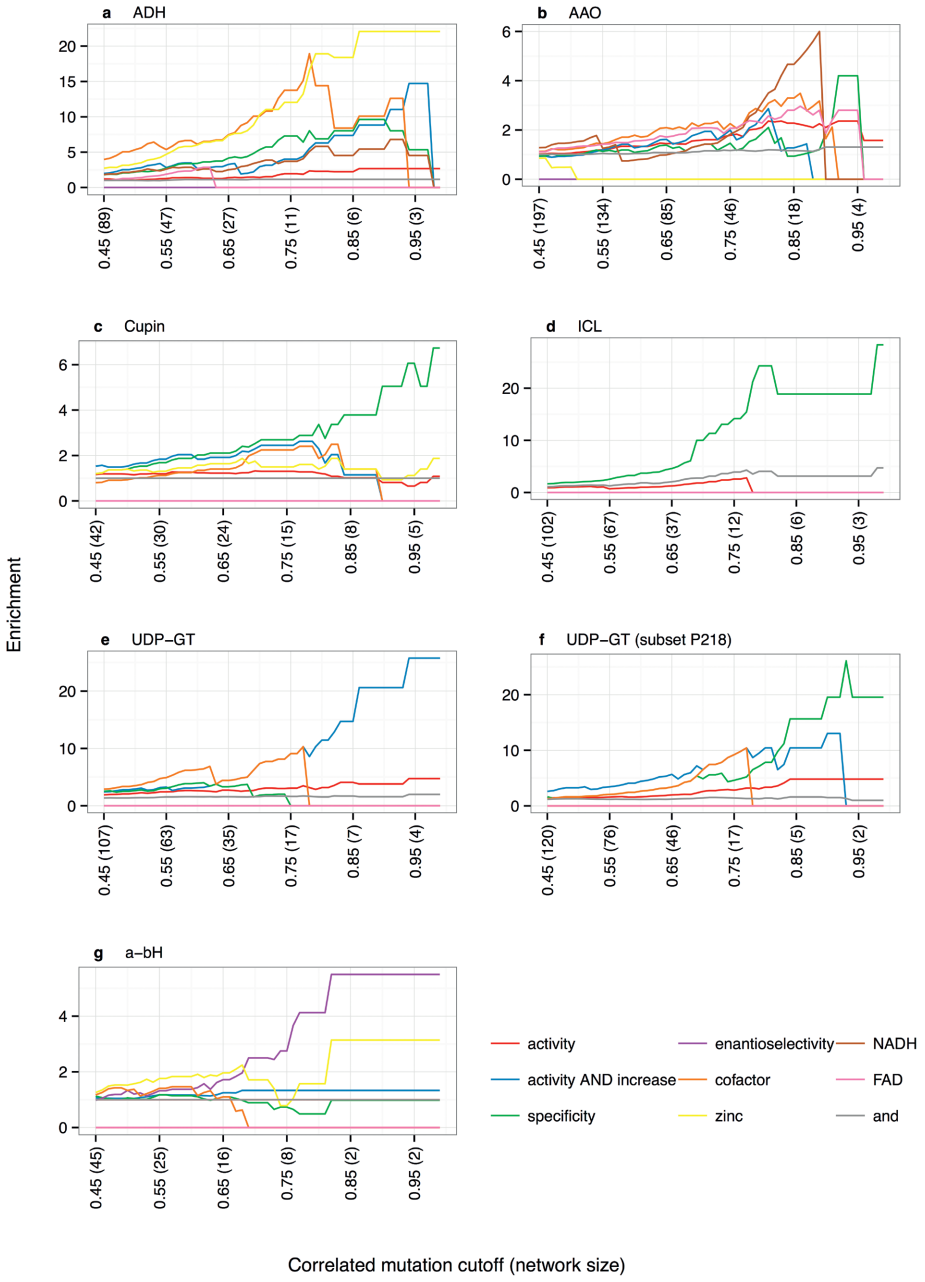


Fig 2. Escores for a series of keywords related to mutations in the families as function of the correlated mutation analysis cut-off. **a.** Keyword enrichments for the alcohol dehydrogenases (ADH). **b.** Keyword enrichments for the Amino acid oxidases (AAO). **c.** Keyword enrichments for the Cupins. **d.** Keyword enrichments for the isocitrate lyases (ICL). **e.** Keyword enrichments for the UDP-Glycosyltransferases (UDP-GT). **f.** Keyword enrichments for a subset of the UDP-Glycosyltransferases (UDP-GT) alignment. This subset is composed of all sequences that have a proline at 3D-number 218. **g.** Keyword enrichments for the a/b-hydrolases (a-bH).

2.4 ADH family

Multiple keywords are enriched for the alcohol dehydrogenase family network (Fig. 2a). At a correlation cut-off of 0.80 most of the positions in the network are located in the active site and many of the residues at these positions will likely have more than one function. The difference between the keyword 'activity' and the joint-keyword 'activity AND increase' should also be noted. Mutations reported in the literature combined with the keyword 'activity' are far more evenly distributed over the alignment positions than mutations combined with the keyword 'activity AND increase', which is much more enriched for alignment positions within the correlation network. This indicates that many of the positions that can be mutated to increase the activity of these proteins are within the network.

2.5 AAO family

The amino acid oxidase Escores show that its Network is mainly enriched for 'FAD', 'co-factor', and 'specificity'. Fig 1c shows that the AAO network consists of two sub-networks; the one surrounding the FAD cofactor (blue positions) and the other surrounding the substrate-binding pocket (red positions). The enrichments shown in Fig 2b are the sum of the two sub-networks. In fact, mutations related to the keywords 'FAD' and 'cofactor' are more abundant in the blue sub-network and mutations related to the keyword 'specificity' are mostly detected in the red sub-network.

2.6 Cupin family

Fig 2c shows enrichment for specificity in correlating positions in the cupin super-family Network. At a low cut-off this network shows a low enrichment for 'activity AND increase', for 'co-factor' and for 'specificity'. In contrast to the AAO correlation network, the cupin Network is not divided into separate sub-networks. However, a closer investigation of the positions leading to these enrichments revealed that the 'specificity' related positions are other positions than the 'cofactor' and the 'activity AND increase' related positions. S1f Fig shows this network in the 3D structure.

2.7 ICL family

In the ICL super-family, very high Escores are observed using 'specificity' as keyword suggesting that specificity is the driving force causing these residues positions to mutate simultaneously. Inspection of the 3D location of this network reveals that the residues are mainly located in and around the active site (Fig. 1a). Escores for the control keyword 'and', also illustrated in Fig 2d, show that this keyword is slightly over-represented in this family. Apparently, the majority of the relatively small number of mutations made in proteins of this family are located at positions surrounding the active site probably due to biased selection of residues by scientists.

2.8 UDP-GT family

The joint-keyword 'activity AND increased' is clearly the enriched in the UDP-GT protein super-family network (Fig. 2e). Note that, like in the ADH family, the keyword 'activity' is hardly enriched in this network. Fig 2f shows that in a subset of the UDP-GT super-family composed by sequences that have a proline at 218, 'specificity' clearly has the highest Escore. This subset is discussed in more detail below. S1e Fig shows both the main network and the network for the subset in the 3D structure.

2.9 a-bH family

In the a/b-hydrolase fold super-family CMA the keyword "enantioselectivity" is clearly enriched (Fig. 2g). The Network consists of two sub-networks and most of the mutations effecting enantioselectivity are located in one of the sub-networks (Fig. 1d). For five of the ten positions of this sub-network, mutations have been published that effected enantioselectivity (shown in cyan in Fig. 1d). To test if the other positions in this network are also important for enantioselectivity a small mutant library was generated for the five non-annotated positions (shown in green in Fig. 1d). The positions of the second sub-network cluster spatially, and are lightly enriched for the keyword 'specificity'.

2.10 Mutant library

The results of an esterase mutation study (Table 2) clearly show the expected impact of the selected correlated network positions on enantioselectivity: 17% of all variants exhibited an improved enantioselectivity (data available in S1 Table) compared to wild-type esterase. Best results were found after the combination of the best mutations obtained at positions 61 (G61S) and 81 (K81H), which led to a 2-3 fold improvement in enantioselectivity.

Table 2. Specific activities and apparent enantioselectivity for the top esterase variants.

Variant	Specific activity[a] [mU/mg]		E_{app} [b]
	(*R*)-3PB-pNP	(*S*)-3PB-pNP	
Wild-type	1.44 (± 0.09)	0.30 (± 0.11)	5
K81H	3.22 (± 0.19)	0.54 (± 0.03)	6
G61S	4.48 (± 0.72)	0.47 (± 0.04)	10
G61S/K81H	6.86 (± 1.08)	0.51 (± 0.03)	13

[a] One unit corresponds to 1 μmol converted min^{-1} mg^{-1} protein.
[b] E_{app} is the ratio of activity for the two enantiomer of (*R*)- and (*S*)-3PB-pNP.

A structural analysis of these two positions revealed that position 61 is in the active site region of the esterase from *Pseudomonas fluorescens* (PFE) adjacent to the catalytic aspartic acid, which suggests that a mutation at this position could influence selectivity[26] although the risk is high that catalytic activity can be strongly affected.

In contrast position 81 is located on the surface of the protein, far away from the active site. Selection of this position without the CorNet tool and 3DM would have been rather unlikely. The increase in the enantioselectivity is clearly cumulative, although the two positions do not correlate directly to each other in the network.

2.11 Co-evolution networks in alignment subsets

Which function is the underlying force behind a CMA network heavily depends on the input alignment. The Networks of large alignments that cover a large evolutionary spread (e.g. a complete super-family) is composed of different positions compared to a Network of subsets of these alignments that cover only a phylogenetic sub-branch of the large alignment. To investigate the effects of selecting sub-branches on the location of CMA networks in the three-dimensional structure several subgroups of the ADH super-family, the ICL super-family, and the UDP-GT super-family were composed. In all three super-families, sub-alignments were generated by selecting a subgroup of sequences that have a residue conserved at the hub of the main Network.

The Network of the ADH super-family alignment, for instance, contains a single network and no clear sub-networks can be detected. This network is located in the centre of the active site (red residues Fig. 3) and surrounds the zinc ion that is essential for the catalytic activity. Position 41 is a hub in this network (Fig. 1b) and makes physical contact with the zinc ion clearly indicated by the high Escore for 'zinc' (Fig. 2a). The two main residues observed at position 41 are Cys (present in 52.4% of the super-family sequences) and Asn (present in 37.8% of the sequences). A sub-alignment was generated using sequences having a cysteine at alignment position 41. In this subset, position 41 is obviously fully conserved and thus no longer shows up in any correlated mutation network. We observe two new networks in this subset (yellow in Fig. 3). These are located surrounding the Network of the complete super-family more in the second layer of the active site. Position 159 is now the main hub in the most extensive network and position 159 mainly occupied by a Gly in the MSA. Using a subset of sequences that have both a Cys at position 41 and a Gly at position 159 we obtain yet another network (blue in Fig. 3) positioned in the third layer around the active site. This sub-location of the Networks in different layers around the active site suggests that they reflect different roles (e.g. activity, specificity, dimerization, etc.) that the corresponding residues need to perform. Unfortunately, for the ADH protein family no literature data is yet available that proves this hypothesis and no function could be assigned to the sub-networks with the available literature data.

The same experiment was performed on the ICL super-family. Position 157 is the main hub of the Network in this super-family and proline is the most common residue at position 157. Fig 2d shows that the main function underlying the Network of this super-family is specificity and this network is located surrounding the substrate-binding pocket (Fig. 1a). Although, also for this family, not enough literature-derived mutation

data are available to prove the function of the residues in the Network that was generated for a subset containing only sequences with a proline at 157, the network is located almost exclusively at the dimerization interface (Fig. 4).

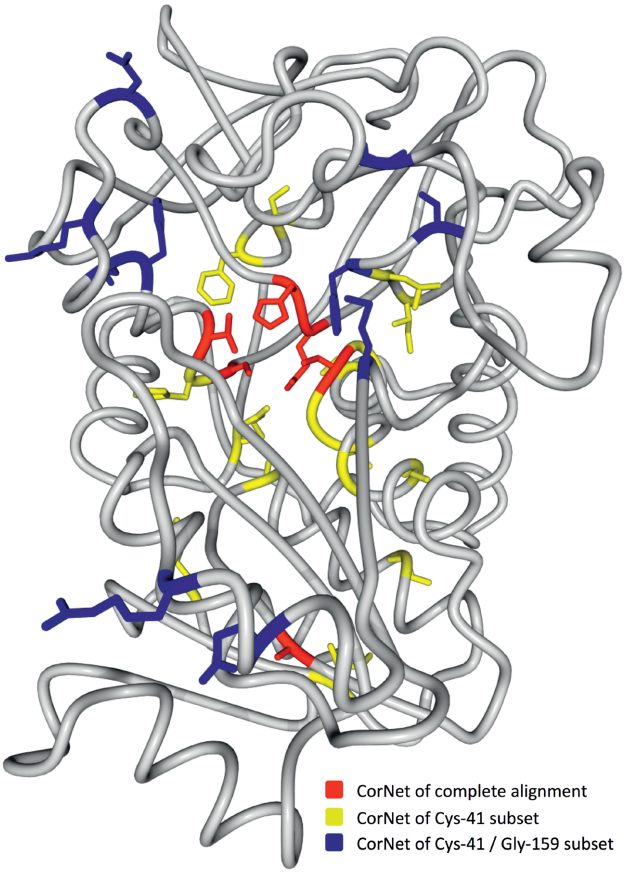


Fig 3. Alcohol dehydrogenase family structure 1CDO-A with CMA network positions of three different alignments visualized. The red residues represent the CMA positions for the complete super-family alignment. The yellow residues represent a network generated for a sub-alignment composed of sequences with a cysteine on 3D-number 41. The blue residues reflect a Network generated for a sub-sub-alignment composed of sequences with a cysteine at position 41 and a glycine at position 159.

This experiment was repeated in the UPD-GT super-family of which the main network shows a high Escore for the keyword "activity AND increase" (Fig. 2e). Position 218 is the main centre of the Network of this super-family and proline is the most common residue at this position. An alignment was generated of all sequences that have a proline at position 218. As shown in Fig 2f in the Network of this alignment the keyword "specificity" results in the highest enrichment.

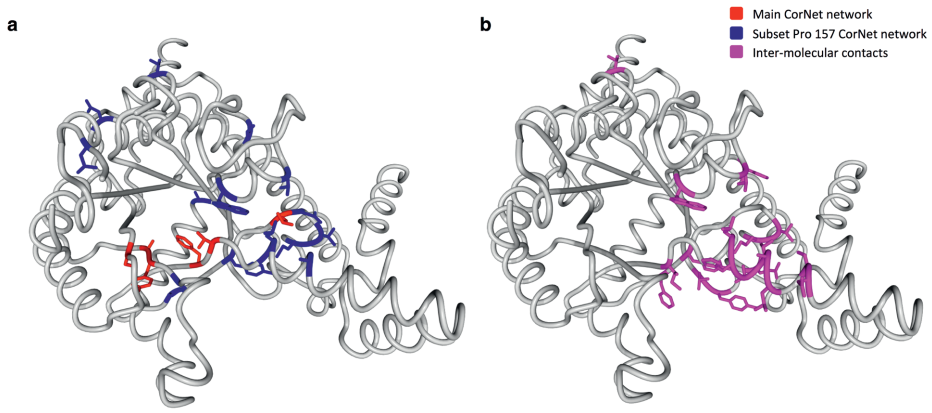


Fig 4. Isocytrate lyases family structure 1DQU-A with CMA networks and dimer interface visualized. **a.** The red residues represent the Network for the complete super-family alignment. The blue residues represent the network for an alignment subset that contains a proline on 3D-number 157. **b.** The purple residues represent the 3D-positions that make an inter-molecular contact in most of the 70 available structures of the ICL family.

2.12 Correlation networks in random subsets

To define the minimal number of sequences needed to perform CMA subsets of randomly selected sequences were generated for all six superfamilies. For each superfamily a range of subsets was generated that contained between 0.5% and 60% of the superfamily sequences. For each subset the network positions were compared to the network of the full alignment and an F-measure was calculated to determine the similarity between the networks. These results (S1 File), show that an alignment of 500 sequences usually contain enough signal to result in a reliable CorNet network indicated by an F-measure of 0.8 or higher.

3 DISCUSSION

We describe the protein function - structure - CMA relations for six protein super-families, which were validated using available mutation data from literature. For one of them, the a-bH protein family, a smart mutant library consisting of only 72 variants was designed based on CorNet network to validate the predicted effect on enantioselectivity. To show that positions in a co-evolution network share a common function this library was targeted only at nodes of the network for which no literature mutation data was available describing effects on enantioselectivity. Although the changes observed in enantioselectivity (2-3 fold) are not very large, nearly 20% of the mutants in the library had an effect on enantioselectivity. Typically, random generated libraries have a hit rate of about 1% (Reetz et. al. [27]). This result clearly shows that the positions in a CorNet network are often functionally related. Therefore, mutation information that is available for nodes in a CorNet network can be used to predict the effects of mutating the non-annotated nodes. This experiment was not performed to create a highly en-

antioselective enzyme (in that case nodes for which effects on enantioselectivity were already published should have been included), but the goal of this experiment was to show that CorNet can be used to find novel mutation hotspots not reported in literature before. In fact, in a recent study we generated a highly selective PFE by mutating a CorNet position for which mutational data was available in the literature [28].

A CMA network, and the function(s) it reflects, depend on the sequences in the alignment. This work shows that not the number of sequences in the alignment, but the evolutionary spread of the aligned sequences is the determining factor for the composition of a CMA network. A large evolutionary spread among the aligned sequences tend to result in a network composed of positions near the active site (i.e. residues performing the main task of the protein). An alignment based on a subset of sequences with a smaller evolutionary spread (i.e. by demanding that one functionally important residue is conserved throughout the subset) results in a correlated mutation network located in the second or third layer of residues (i.e. residues involved in more specific functions). This phenomenon was nicely demonstrated by the difference of enrichments scores in the UDP-GT protein family, where in the full alignment "activity AND increase" resulted in the highest Escore whereas "specificity" scored highest in much smaller subset of the alignment (where P218 is conserved). Rules for determining the best set of input sequences that will result in a Network optimized for a specific protein feature, still remains to be determined. The alignments used in this work were, in fact, automatically generated and no filtering or any form of optimizing was conducted. This shows not only that this method is robust but also that there is still much room for further developments, improvements, and novel discoveries in the area of CMA network related research. The fact that the maximum Escores differ between different super-families and for different protein features suggests that the alignments, and especially the selection of sequences to be included, can be optimized even further. The accuracy of the type of analysis conducted is this work increases when more data is available for a super-family as indicated by high Escores of control keywords in the rather small ICL super-family. Together with the explosion of sequence- and mutation data that is becoming readily available we believe that the type of protein data analysis described in this work might become a standard tool for protein engineering.

4 MATERIALS AND METHODS

4.1 Protein families

The relation between correlation networks and information from the bibliome was analysed for six super-families. *ADH:* Alcohol dehydrogenases catalyse the oxidation of alcohols by the reduction of nicotinamide adenine dinucleotide. *AAO:* Amino acid oxidases are FAD-binding proteins. This family consists of two sub-families that catalyse *D*-amino acid and *L*-amino acid conversion, respectively. These two AAO sub-families bind their FAD differently. *Cupin:* The very large RmlC-like cupin family comprises a wide range

of enzymes that can convert many different substrates. Cupins show a large variety of reaction mechanisms. The cupins are the most diverse protein family known today covering 17 enzyme classes and even other types of proteins such as seed storage globulins and multi-domain transcription factors[29]. *ICL:* The phosphoenolpyruvate mutase/isocitrate lyases super-family contains several enzyme families that act on alpha-oxycarboxylate substrates. *UDP-GT:* The UDP-Glycosyltransferases protein family contains sugar-acting enzymes that can act on different sugars and perform different reactions (synthases, transferases, phosphorylases). *a-bH:* The a/b-hydrolase fold super-family contains a wide range of proteins including proteases, esterases and lipases[30].

For each of these six families structure based MSAs were produced, and the literature was scanned for mutations. Table 3 lists the number of articles, sequences, structures, core alignment positions, and mutations found for each of the six protein super-families.

Table 3. Sequences, structures, and mutations found for the six super-families.

Name	Sequences	Core alignment positions	Structures	Articles scanned	Mutation data extracted
ADH	14696	353	447	15144	10437
AAO	12155	253	356	14442	6203
Cupin	1650	43	338	53400	4362
ICL	3019	170	70	2013	160
UDP-GT	36402	313	475	26919	7610
a-bH	59904	88	1665	60926	60755

The CMA scores were determined for all pairs of alignment positions in each of the six families. Mutual information was calculated rather than the direct information that has been described[9–12]. Correlation scores are obtained using the previously described Comulator software[14]. Comulator uses a method known as a statistical coupling analysis[31,32] to assign correlation scores. Comulator was used because this method is a robust CMA algorithm that was specifically developed to handle large structure based superfamily alignments that consist of thousands of proteins and often contain many different protein functions.

4.2 Cornet Features

The CMA network visualization tool was built using cytoscape.js (a JavaScript graph visualization library)[33] and jquery (user interface libraries). In this HTML based network viewer nodes represent alignment positions (with the MSA position number indicated) and edges are coloured as function of the pairwise CMA values. The nodes are hyperlinked to underlying data stored in the database so that, for example, the amino acid distribution of an alignment position or a pair of correlating positions can be obtained rapidly.

The user can interactively choose correlation cut-offs, colours for groups of residue positions, and residue positions can be coloured as function of their annotation. CorNet can write the resulting colours in a YASARA scene so that results can be visualised with YASARA, a protein structure visualisation tool (Fig. 1). When annotation queries are performed, the enrichment of the search term is determined on the fly.

4.3 Mutation extraction

The Mutator software[24] was used for the extraction of mutations from the literature. This software searches in PubMed with words like 'mutation', 'SNP' (Single Nucleotide Polymorphism), 'substitution', or 'recombinant' combined with family-specific keywords (and their synonyms) like names of family members, their gene names, names of diseases known to be related to members of the family, or generic family names. Most names are retrieved from the Swiss-Prot[34] protein entries available in the MSA. We subsequently scan these articles for mutation information related to the six super-families. Each sentence that contained a mutation (i.e. terms like S127P, Glu422Lys or "Trp58 was mutated to Ala") was analysed for the presence of a series of words such as 'specificity', 'activity', 'cofactor', etc. A residue position is considered related to a keyword if the combination of mutation and keyword is observed in the same sentence in at least two articles that describe a mutation at the same residue position in a member of the family.

4.4 3DM

3DM was used to generate the structure bases multiple sequence alignments (MSAs) for the six super-families[35]. In summary, structures are superposed with WHAT IF [21] to generate an initial alignment that is then used to guide the alignment of all sequences for which no structure data is available. 3DM allows for the generation of alignments for subgroups of sequences. Such sub-alignments were generated for the UPD- and ADH protein families. These sub-alignments are composed of all sequences that have the most abundant residue at the position that forms the centre of the Network. Correlated mutation analysis is performed as described before[14]. A method known as statistical coupling analysis[32] is used to detect pairs of residues in the alignment that mutate simultaneously.

4.5 3D numbers

CorNet uses a super-family specific residue numbering scheme for all sequences and structures in the alignment. Structurally equivalent residues get the same numbers, called 3D-numbers, which are also used for the corresponding sequence alignment positions. 3D-numbering schemes are used throughout this paper, and in the interactive version of CorNet. The principle of 3D- numbers and the underlying structure based multiple sequence alignment have been described[35] and is illustrated in Fig 5. Structurally variable sites such as residues that reside in loops are not included in the structure based MSA and thus are not included in the correlation analyses. In practice, though, functionally important residues normally are located in the structurally conserved regions of proteins.

Fig 5. Example to illustrate the use of 3D-numbers. We are interested in histidine 22 in the human sequence, however mutation related information from the bibliome is only available for the mouse homologous sequence. In the main text we find a description of the effect of a mutation of histidine 49 to an alanine. This histidine residue is in the structure at equivalent position of the human histidine-22 and therefore shares the same 3D number (17).

4.6 Escore and P-values

The keyword search option enables the user to automatically select mutations for which that keyword is part of the annotation and to map these on the network. The overrepresentation of a keyword for residues in a network is expressed as the enrichment-score (Escore).

$$Escore = (K_n \,/\, K_t) \ * \ (N_t \,/\, N_n)$$

Equation 1. N_n = number of alignment positions in the network, N_t = total number of alignment positions, K_n = number of network positions for which the keyword m was observed, K_t = total number of positions for which keyword m was observed.

CorNet offers the user to define a cut-off (N_{mut}), which can be selected interactively; the default is 2. The keyword must be observed with the same 3D residue position in at least N_{mut} mutation studies for different proteins in order to be accepted.

4.7 Library design

The design of the mutant library composed of 72 variants was based on a 3DM analysis of the respective positions, which led to the incorporation of the four most frequent amino acids at the networks positions (Table 4): a triple mutant library was designed to include the combinatorial effects of those positions that either are connected with more than one node with a known effect on enantioselectivity (i.e. nodes 27 and 61, Fig. 1d) or with nodes that have been more frequently mutated according to literature (i.e. node 14, Fig. 1d). The remaining two positions (i.e. nodes 59 and 81) were randomized independently.

Table 4. 3D positions selected, codons used for library design and corresponding encoded amino acids.

3D position	Codons	Amino acids encoded
14	TKG/TWT	L,W,**F**,Y
27	GBC/ACC	V,A,G,**T**
59	VTT/GGT	V,**I**,L,G
61	GSC/ARC	**G**,A,N,S
81	YAT/CGT/GTT	H,Y,R,V

Residues in bold correspond to wild-type esterase.

4.8 Mutant libraries and enantioselectivity

Libraries of the esterase from *Pseudomonas fluorescens* (PFE) were constructed by QuikChange mutagenesis. In the case of the triple mutant library three consecutive reactions were needed. In each case the following reaction mixture was prepared: sterilized deionized H_2O (41 µL), Pfu buffer (10x, 5 µL), dNTP (1 µL, 10 mM each), plasmid pJOE2792.1 (1 µL, 50 nmol μL^{-1}) containing the gene encoding PFE[36], mixture of forward and reverse primer mixture (1 µL, 12.5 nmol μL^{-1}), Pfu^+ DNA Polymerase (0.2 µL). The so prepared mixture was then split in two different PCR tubes with equal amount of volumes and used for a PCR at the following conditions: 1) 95°C, 300 s; 2) 30 cycles: 95°C, 30 s; 50 or 65°C, 30 s; 72°C 210 s; 3) 72°C, 480 s. Afterwards, the presence of the PCR product was verified on a 1% agarose gel and finally DpnI (0.5 µl) was added to remove the template. Digestion of the most abundant product was performed for 2 h at 37°C followed by denaturation of DpnI at 80°C for 20 minutes. Chemo-competent *E. coli* cells (Top10) were transformed with the PCR product for plasmid amplification and quality library evaluation[37]. Once the randomization state of the mutated position was verified by sequencing, the mixture of circularized plasmids was used for transformation in chemo-competent *E.coli* cells (BL21 DE3) and plated onto LB_{AMP}-plates. Clones were picked with sufficient oversampling (3-fold) to ensure statistically a 95% coverage of the library[38].

4.9 Primers:

1afw - 5′-GGTGTTGTKGAGCCACGGTTGGCTACTGG-3′,
1bfw - 5′-GGTGTTGTWTAGCCACGGTTGGCTACTGG-3′,
1arv - 5′-CGTGGCTCMACAACACCGGTTTACCGCTGC-3′,
1brv - 5′CGTGGCTAWACAACACCGGTTTACCGCTGC-3′,
2afw - 5′-CCTCAAGGAGGTGGBCCTGGTGGGCTTCTCC-3′,
2bfw - 5′-CCTCAAGGAGGTGACCCTGGTGGGCTTCTCC-3′,
2arv - 5′-GGAGAAGCCCACCAGGVCCACCTCCTTGAGG-3′,
2brv - 5′-GGAGAAGCCCACCAGGGTCACCTCCTTGAGG-3′,
3afw - 5′- CCACCCTGGTGVTTCATGGCGATGG-3′,
3bfw - 5′- CCACCCTGGTGGGTCATGGCGATGG-3′,

3arv - 5′-CCATCGCCATGAABCACCAGGGTGG-3′,
3brv - 5′- CCATCGCCATGACCCACCAGGGTGG-3′,
4afw - 5′-GTGATCCATGSCGATGGCGACC-3′,
4bfw - 5′- GTGATCCATARCGATGGCGACC-3′,
4arv - 5′- GGTCGCCATCGSCATGGATCAC-3′,
4brv - 5′- GGTCGCCATCGYTATGGATCAC-3′,
5afw - 5′- CGAACTGYATGTGTACAAGGACG-3′,
5bfw - 5′- CGAACTGCGTGTGTACAAGGACG:-3′,
5cfw - 5′- CGAACTGGTTGTGTACAAGGACG-3′,
5arv - 5′- CGTCCTTGTACACATRCAGTTCG-3′,
5brv - 5′- CGTCCTTGTACACACGCAGTTCG-3′,
5crv - 5′- CGTCCTTGTACACAACCAGTTCG-3′,
6afw - 5′- GCCGAACTGCATGTGTACAAGGACGCGCCCCACG-3′,
6arv - 5′- CCTTGTACACATGCAGTTCGGCGCCCTTGATCAAC-3′,
7afw - 5′- GGTGGTGCATAGCGATGGCGACCAGATCG-3′**,**
8arv - 5′- CGCTATGCACCACCAGGGTGGGTACGTC-3′.

The primers series **1**, **2** and **4** were used for the randomization of positions 14, 27 and 61, respectively, in the triple mutant library. Primers series **3** and **5** were used for the independent randomizations at positions 59 and 81 respectively. Primers series **6** and **7** were used for the creation of the single mutants derived from the combination of the best hits at each network node.

For protein expression, the transformants were grown on agar plates, picked and inoculated into microtiter plates containing 200 µL LB_{AMP}. Incubation was performed overnight at 37°C and 500 rpm. The following day the overnight culture (50 µL) was transferred into deep-well blocks containing 1 mL TB_{AMP} and incubated for 3 h at 37°C at 700 rpm. Gene expression was induced with L-rhamnose solution (final concentration 0.2% (w/v)). The libraries were incubated for an additional 16 h at 30°C, 700 rpm. For disruption, cells were harvested by centrifugation (15 min, 4355 g and 4°C) and resuspended in 300 µL lysis buffer containing 1% Bugbuster solution for 1 h at 37°C at 700 rpm followed by centrifugation for 45 min at 4355 g, 4°C. The crude cell extract was transferred into a new microtiter plate and stored until usage at 4°C. For each variant the crude cell lysate was split into two microtiter plates containing phosphate buffer (50 mM, pH 7.5). Enantioselectivity measurements were performed in microtiter plates (MTP) first with crude cell lysate using optically pure (*R*)- and (*S*)-3-phenylbutyric acid-*p*-nitrophenylesters (0.2 mM final concentration in 20 % acetonitrile, synthesized as described previously[39]) in two separate wells of the MTP for each variant following for 1 h the increase in absorbance at 410 nm from the released *p*-nitrophenolate. From the difference in the rate of the hydrolysis of the two enantiomers, the apparent enantioselectivity was determined as described previously[39]. Variants showing improved

properties in this initial screening were produced on larger scale, His-tag purified using TALON beads and reanalyzed for altered enantioselectivity.

SUPPORTING INFORMATION

Supplementary data is available online:
https://figshare.com/s/6f19c20c1bb29de134c6

AUTHOR CONTRIBUTIONS

Conceptualization: TB HJ RJ BV. Data curation: TB RK. Formal analysis: TB. Funding acquisition: HJ UB. Investigation: TB KS AN TD YT TT RK. Methodology: TB GT BV. Project administration: HJ. Resources: HJ YT TT UB TD BN. Supervision: HJ PS UB GV. Validation: TB AN YT. Visualization: TB. Writing – original draft: TB GT. Writing – review & editing: TB GT AN.

REFERENCES

1. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, et al. Environmental genome shotgun sequencing of the Sargasso Sea. Science. 2004;304: 66–74. doi:10.1126/science.1093857
2. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, et al. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol. 2007;5: e77. doi:10.1371/journal.pbio.0050077
3. Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, et al. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol. 2007;5: e16. doi:10.1371/journal.pbio.0050016
4. Bornscheuer UT, Huisman GW, Kazlauskas RJ, Lutz S, Moore JC, Robins K. Engineering the third wave of biocatalysis. Nature. 2012;485: 185–194. doi:10.1038/nature11117
5. Kowarsch A, Fuchs A, Frishman D, Pagel P. Correlated Mutations: A Hallmark of Phenotypic Amino Acid Substitutions. PLoS Comput Biol. 2010;6: e1000923. doi:10.1371/journal.pcbi.1000923
6. Oliveira L, Paiva ACM, Vriend G. Correlated mutation analyses on very large sequence families. ChemBioChem. 2002;3: 1010–1017. doi:10.1002/1439-7633(20021004)3:10<1010::AID-CBIC1010>3.0.CO;2-T
7. Oliveira L, Paiva PB, Paiva ACM, Vriend G. Identification of functionally conserved residues with the use of entropy-variability plots. Proteins. 2003;52: 544–552. doi:10.1002/prot.10490
8. Moult J, Pedersen JT, Judson R, Fidelis K. A large-scale experiment to assess protein structure prediction methods. Proteins. 1995;23: ii–v. doi:10.1002/prot.340230303
9. Burger L, van Nimwegen E. Disentangling direct from indirect co-evolution of residues in protein alignments. PLoS Comput Biol. 2010;6: e1000633. doi:10.1371/journal.pcbi.1000633
10. Jones DT, Buchan DWA, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics. 2012;28: 184–190. doi:10.1093/bioinformatics/btr638
11. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, et al. Protein 3D structure computed from evolutionary sequence variation. PLoS ONE. 2011;6: e28766. doi:10.1371/journal.pone.0028766
12. Marks DS, Hopf TA, Sander C. Protein structure prediction from sequence variation. Nat Biotechnol. 2012;30: 1072–1080. doi:10.1038/nbt.2419
13. Folkertsma S, van Noort P, Van Durme J, Joosten H-J, Bettler E, Fleuren W, et al. A family-based approach reveals the function of residues in the nuclear receptor ligand-binding domain. J Mol Biol. 2004;341: 321–335. doi:10.1016/j.jmb.2004.05.075
14. Kuipers RKP, Joosten H-J, Verwiel E, Paans S, Akerboom J, van der Oost J, et al. Correlated mutation analyses on super-family alignments reveal functionally important residues. Proteins. 2009;76: 608–616. doi:10.1002/prot.22374
15. Leferink NGH, Fraaije MW, Joosten H-J, Schaap PJ, Mattevi A, van Berkel WJH. Identification of a gatekeeper residue that prevents dehydrogenases from acting as oxidases. J Biol Chem. 2009;284: 4392–4397. doi:10.1074/jbc.M808202200
16. Wilkins AD, Bachman BJ, Erdin S, Lichtarge O. The use of evolutionary patterns in protein annotation. Curr Opin Struct Biol. 2012;22: 316–325. doi:10.1016/j.sbi.2012.05.001

17. Proctor EA, Kota P, Demarest SJ, Caravella JA, Dokholyan NV. Highly covarying residues have a functional role in antibody constant domains. Proteins. 2013;81: 884–895. doi:10.1002/prot.24247

18. Sreekumar J, ter Braak CJF, van Ham RCHJ, van Dijk ADJ. Correlated mutations via regularized multinomial regression. BMC Bioinformatics. 2011;12: 444. doi:10.1186/1471-2105-12-444

19. Gouldson PR, Dean MK, Snell CR, Bywater RP, Gkoutos G, Reynolds CA. Lipid-facing correlated mutations and dimerization in G-protein coupled receptors. Protein Eng. 2001;14: 759–767.

20. Göbel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins. 1994;18: 309-317. doi:10.1002/prot.340180402

21. Vriend G. WHAT IF: a molecular modeling and drug design program. J Mol Graph. 1990;8: 52–56, 29.

22. Marengere LE, Pawson T. Structure and function of SH2 domains. J Cell Sci Suppl. 1994;18: 97–104.

23. Halabi N, Rivoire O, Leibler S, Ranganathan R. Protein sectors: evolutionary units of three-dimensional structure. Cell. 2009;138: 774–786. doi:10.1016/j.cell.2009.07.038

24. Kuipers RK, Bergh T van den, Joosten H-J, Lekanne dit Deprez RH, Mannens MM, Schaap PJ. Novel tools for extraction and validation of disease-related mutations applied to Fabry disease. Hum Mutat. 2010;31: 1026–1032. doi:10.1002/humu.21317

25. Vroling B, Thorne D, McDermott P, Attwood TK, Vriend G, Pettifer S. Integrating GPCR-specific information with full text articles. BMC Bioinformatics. 2011;12: 362. doi:10.1186/1471-2105-12-362

26. Park S, Morley KL, Horsman GP, Holmquist M, Hult K, Kazlauskas RJ. Focusing mutations into the P. fluorescens esterase binding site increases enantioselectivity more effectively than distant mutations. Chem Biol. 2005;12: 45–54. doi:10.1016/j.chembiol.2004.10.012

27. Reetz MT, Zonta A, Schimossek K, Jaeger K-E, Liebeton K. Creation of Enantioselective Biocatalysts for Organic Chemistry by In Vitro Evolution. Angew Chem Int Ed Engl. 1997;36: 2830–2832. doi:10.1002/anie.199728301

28. Nobili A, Tao Y, Pavlidis IV, van den Bergh T, Joosten H-J, Tan T, et al. Simultaneous use of in silico design and a correlated mutation network as a tool to efficiently guide enzyme engineering. Chembiochem. 2015;16: 805–810. doi:10.1002/cbic.201402665

29. Dunwell JM, Purvis A, Khuri S. Cupins: the most functionally diverse protein superfamily? Phytochemistry. 2004;65: 7–17.

30. Kourist R, Jochens H, Bartsch S, Kuipers R, Padhi SK, Gall M, et al. The alpha/beta-hydrolase fold 3DM database (ABHDB) as a tool for protein engineering. ChemBioChem. 2010;11: 1635–1643. doi:10.1002/cbic.201000213

31. Fodor AA, Aldrich RW. On Evolutionary Conservation of Thermodynamic Coupling in Proteins. J Biol Chem. 2004;279: 19046–19050. doi:10.1074/jbc.M402560200

32. Lockless SW, Ranganathan R. Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. Science. 1999;286: 295–299. doi:10.1126/science.286.5438.295

33. Lopes CT, Franz M, Kazi F, Donaldson SL, Morris Q, Bader GD. Cytoscape Web: an interactive web-based network browser. Bioinformatics. 2010; doi:10.1093/bioinformatics/btq430

34. Magrane M, Consortium U. UniProt Knowledgebase: a hub of integrated protein data. Database. 2011;2011: bar009-bar009. doi:10.1093/database/bar009

35. Kuipers RK, Joosten H-J, van Berkel WJH, Leferink NGH, Rooijen E, Ittmann E, et al. 3DM: systematic analysis of heterogeneous superfamily data to discover protein functionalities. Proteins. 2010;78: 2101–2113. doi:10.1002/prot.22725

36. Krebsfänger N, Zocher F, Altenbuchner J, Bornscheuer UT. Characterization and enantioselectivity of a recombinant esterase from Pseudomonas fluorescens. Enzyme and Microbial Technology. 1998;22: 641–646. doi:10.1016/S0141-0229(98)00004-0

37. Kille S, Acevedo-Rocha CG, Parra LP, Zhang Z-G, Opperman DJ, Reetz MT, et al. Reducing codon redundancy and screening effort of combinatorial protein libraries created by saturation mutagenesis. ACS Synth Biol. 2013;2: 83–92. doi:10.1021/sb300037w

38. Reetz MT, Kahakeaw D, Lohmer R. Addressing the numbers problem in directed evolution. ChemBioChem. 2008;9: 1797–1804. doi:10.1002/cbic.200800298

39. Jochens H, Bornscheuer UT. Natural diversity to guide focused directed evolution. ChemBioChem. 2010;11: 1861–1866. doi:10.1002/cbic.201000284

CHAPTER 5

Inherited arrhythmia syndromes, how to identify pathogenic mutations?

Bas Vroling*, Tom van den Bergh*, Marielle Alders, Stephan Heijl, Michael Tanck, Ronald Lekanne Dit Deprez, Remko Kuipers, Jamie D. Kapplinger, Michael Ackerman, Henk-Jan Joosten, Arthur Wilde#, Marcel Mannens#

** Authors contributed equally to this manuscript*

 # Authors contributed equally to this manuscript

TO BE SUBMITTED

ABSTRACT

Background Long QT Syndrome (LQTS) and Brugada Syndrome (BrS) are cardiac arrhythmia syndromes associated with sudden cardiac death. In particular in LQTS, genetic testing is pivotal for the diagnosis, prognosis and treatment of patients. Due to the lack of evidence of pathogenicity generated by functional analysis, improved *in silico* systems for the analyses of these missense variants are therefore of critical importance

Methods We evaluated a total of 1,238 known variants. Of these 1,238 variants, 719 variants were associated with a LQTS or BrS phenotype, whereas 519 variants were observed in healthy individuals. The variants were observed in KNCH2, KCNQ1 and SCN5A, the three genes responsible for most of the LQTS and BrS phenotypes. We used 3DM to generate many heterogeneous data to describe each variant. Using this data, we generated a pathogenicity predictor (3DMPP) to separate disease-causing from non-disease causing variants.

Results We show that 3DMPP outperforms frequently used genome-wide prediction tools and even gene-family specific predictors, and that these existing tools are not able to accurately assess the pathogenicity of variants.

Conclusions For sudden death–predisposing diseases such as LQTS and BrS accurate and proper interpretation of the genetic test is of paramount importance considering the diagnostic, prognostic, and therapeutic implications of a genetic test. The accurate predictions made by 3DMPP shows that the availability of reliably annotated patient cohort data together with the use of high-quality, domain-specific structure based integrated data is essential for creating a reliable prediction tool.

1 INTRODUCTION

Identification of pathogenic alterations in the DNA sequence of patients with inherited diseases is one of the main tasks of human genetics. Many of putative disease-related variants are located in protein-coding regions of the genome, affecting the structure and function of the encoded protein. More than a hundred thousand unique disease causing mutations are available in the Human Gene Mutation Database (HGMD)[1] that are known to cause over ten thousand different monogenic disorders, and almost four thousand genes already have been described that are involved in polygenic disorders[2]. Next generation gene sequencing efforts have, on the other hand, revealed that harmless single nucleotide polymorphisms (SNPs) are even more frequent in the human genome. A recent study revealed more than 3.5 million alterations in the whole genome of a single individual[3]. Obviously, only a small fraction of these SNPs affect protein function and an even smaller fraction of those variants can be classified as pathogenic.

Whereas mutations predicted to result in truncated proteins (nonsense, frameshift and most splice site mutations) are likely to affect protein function and to be pathogenic, predicting the pathogenicity of missense mutations is much more difficult. Missense variants can be classified as pathogenic in case linkage analysis in families is available (with the exception of linkage disequilibrium). Evidence for pathogenicity generated by functional analysis is helpful but in the field of ion channels notable exceptions exist[4]. Moreover, these data often lack completely and the pathogenicity of a variant remains uncertain. In a diagnostic setting this means that in many cases no clear message can be communicated to the patient and no presymptomatic testing in family members can be offered. Improved *in silico* systems for the analyses of these missense variants are therefore of critical importance. With the availability of next-generation sequencing techniques with massive data, including many missense variants, ensuing, a structured reliable variant screening tool is of paramount importance.

Long QT syndrome (LQTS)

LQTS is one of the cardiac arrhythmia syndromes associated with sudden cardiac death. To date 14 genes have been recognised to cause LQT syndrome. By far, most putative causal variants are identified in genes KCNQ1, KCNH2 and SCN5A. The first 2 genes encode the alpha subunits of two potassium channels and SCN5A encodes the alpha subunit of the cardiac sodium channel. Loss-of-function variants in the potassium channel genes and gain-of-function variants in SCN5A result in net decreased repolarizing force, with subsequent prolongation of the cardiac action potential and the QT interval on the ECG. Brugada syndrome (BrS), among others also caused by variants (loss-of-function) in SCN5A is characterised by ST segment elevation in the right precordial leads (V1-V3) and risk of ventricular fibrillation and sudden cardiac death.

In particular in LQTS, genetic testing is pivotal for diagnosis, prognosis and treatment of patients[5,6].
In this study we used variants found in these genes in LQT patients or BRS patients.
We have previously described 3DM[7], a system that can generate superfamily systems that fully integrate sequence data with literature data, mutation information, and three-dimensional structures. In this study we have used 3DM to generate superfamily information systems for three genes involved in LQTS (KCNH2, KCNQ1, SCN5A) and BrS (SCN5A). Based on the information in these systems we have developed 3DMPP, a tool that can accurately predict the pathogenicity of variants in these arrhythmia-syndrome related genes. We show that 3DMPP clearly outperforms available tools in separating benign variants from pathogenic variants.

2 METHODS

2.1 Dataset

Variant data was collected for three genes: KCNH2, KCNQ1 and SCN5A. Putative pathogenic mutations found in LQT cases include the mutations in the cohorts published by Kapa et al, the cohort of Familion described by Kapplinger et al, mutations detected in the patient population from the Academic Medical Centre (AMC) that were not published before and mutations detected in an additional cohort of Familion that were not published by Kapplinger *et. al*. Mutations in SCN5A found in Brugada syndrome patients include those published by Kapplinger et al and the unpublished mutations found in the AMC cohort. Variants found in the normal population include variants published by Kapa *et. al*. and the variants listed in the Exome Aggregation Consortium (ExAC) database[8].

The dataset contains variants that were concluded to be benign or causal agents of either LQTS or BRS by a variety of methods. Variants for which the effect could not be determined with a high confidence were excluded in this study, since the inclusion of miscategorised variants would impair any generated model. When, in a single patient, multiple variants were found in one of the three genes these were excluded because in these cases it is impossible to address the pathogenicity of the individual mutations.

Variants that were observed in ExAC with a minor allele frequency (MAF) greater than the estimated natural prevalence of LQTS (0.05% or 1 in 2000) were determined as benign. Variants observed with a MAF <0.05% that had not been identified in patients were also classified as benign. However, variants that were observed in the ExAC database only once or had an abnormal functional phenotype were excluded from this study.

Table 1 shows an overview of the training data; each sample was classified with the binary “case” or “control”. Cases are variants linked to LQTS or BRS, whereas controls are benign variants.

Table 1. The distribution of case and control variants through the entire dataset.

	KCNQ1	KCNH2	SCN5A	Total
Cases	183	284	252	719
Controls	88	110	321	519

2.2 3DM information systems

The 3DM software that generates superfamily information systems is extensively described elsewhere[7,9], and only deviations from the normal 3DM procedure will be described here. Traditionally, a structure based multiple sequence alignment (MSA) forms the backbone of the 3DM information systems. However, for the LQT related proteins structure information is available for only very small parts of the proteins. Therefore, a different strategy for creating high-quality alignments for KCNQ1, KCNH2 and SCN5A was used. Since proteins are often composed of distinct structural and functional units, subject to different evolutionary pressures and constraints, separate alignments were created for all domains present in the three individual genes. Sequences included in the alignment were retrieved from UniProt, together with sequences generated based on the variation found in the 1000 genomes project[10] and the NHLBIO GO Exome Sequencing Project. All sequences and structures are renumbered so that residues aligned in the MSA get the same number throughout the information system. This enables the transfer of data and knowledge between proteins, and facilitates literature searches for mutations in homologs. Mutations were extracted from the literature by the 3DM Mutator module[9]. PubMed was queried for papers containing mutations related to proteins present in the LQT-related 3DM information system. Information about mutations in these proteins was subsequently extracted from 111,581 full-text articles. In total, this resulted in 32,874 mutations for the LQT-related protein superfamily.

2.3 Learning features

A total of 41 parameters (shown in supplementary table S1) that describe the characteristics of the missense variants were used as learning features. Each parameter belongs to one of four different categories.

1. Residue-specific parameters

Parameters based on the different inherent characteristics of residue types are likely to determine a part of the impact of the variant. For instance, swapping a positively charged residue with a negatively charged residue might have a larger impact than a neutral change.

2. Alignment-derived parameters

A lot can be learned about the importance and functions of residues by analysing multiple sequence alignments. Attributes such as conservation of the wild type and variant residues, when alignment positions tend to start mutating, and the involvement

of the affected position in correlated mutation networks have been shows to be good indicators of the importance of positions in the protein family[11–14].

3. Structure-based parameters
For those parts of the proteins for which structure information is available, a number of structure-based parameters are calculated. All available structures are superposed in the 3DM information systems, and based on this superposition the structural conservation of positions is determined. In addition, information about the structural position is used, such as whether the mutations reside in a transmembrane or intramembrane segment.

4. Literature-based parameters
For many positions there are papers available describing the function of residues at that position, as well as the roles these residues play in disease when mutated. These studies are available for residues in human proteins as well as for residues in many other species. Because mutations in homologous proteins at the same structural position tend to have similar effects, the information found in the literature for mutations in other species is very valuable, as it can be transferred to the affected human proteins.

For mutations that we find in the literature for human genes we simply check if these mutations are mentioned together with disease terms or not. For mutations described in the literature for homologous proteins, we have chosen a slightly different approach, since most often these mutations are likely to be created by mutagenesis experiments. Based on the assumption that these mutated residues and mutations are more important if they are more frequently discussed in scientific papers (in general, mutations without measurable effects are less often published) we have included a number of parameters that describe the frequencies and types of occurrences of residues and mutations in the literature for non-human proteins.

2.4 Statistical analyses

The Caret R library was used to tune and train models with the data[15]. A 10-fold cross validation approach was used to assess performance. While explorative modelling was done with a variety of algorithms, Random Forest[16], Gradient Boosting Model[17] and Multilayer Perceptron algorithms were initially selected as they consistently performed well. The XGBoost gradient boosting model outperformed the others and was chosen for its performance. To optimize for training and prediction efficiency the final model was built using the R XGBoost package.

3 RESULTS

3.1 Classifier performance

The classifier was trained on the dataset, together with the parameters that were generated by 3DM. The Machine learning performance comparison is based on the Matthew's Correlation Coefficient (MCC)[18]. This score combines accuracy, precision and recall to get a score that allows for an accurate comparison between predictors while it provides a better indication of the actual performance of a predictor than solely the accuracy. An overview of performance metrics for the optimized classifier is given in table 2.

Table 2. 3DMPP classifier performance

Machine	Accuracy	Sensitivity	Specificity	Pos. Pred. Value	Neg. Pred. Value	MCC
3DMPP	0.8113	0.8260	0.7905	0.8408	0.7722	0.6196

3.2 Comparison with other methods

We compared the performance of **3DMPP** with that of a number of existing genome-wide nsSNP-effect predictors and one domain-specific nsSNP effect predictor:

- **S**orting **I**ntolerant **F**rom **T**olerant (**SIFT**): predictions based on the degree of conservation of amino acid residues in sequence alignments[19].
- **PolyPhen-2**: predictions based on physical and comparative considerations[20].
- **Pro**tein **V**ariation **E**ffect **An**alyzer (**PROVEAN**): predictions based on the degree of conservation of amino acid residues in sequence alignments[21].
- **KvSNP**: A machine-learning based predictor optimized specifically for Kv-channel genes[22].
- **V**ariant **E**ffect **S**coring **T**ool (**VEST3**): A genome wide machine-learning based predictor [23].
- **MutationTaster2:** Uses a Bayes classifier to generate predictions[24]**.**
- **MutationAssessor**: A machine learned based predictor that evolutionary conservation patterns[25].
- **LRT**: Model that uses a likelihood ratio test[26].

Results were retrieved from web servers using the default parameters as suggested by the individual tools. Detailed classifier performance statistics are listed in table 3 and a graphical overview of the performance of all predictors is shown in figure 1.

Table 3. Comparison of the performance of the different nsSNP effect predictors. The table is sorted on the Matthews correlation coefficient, a good overall performance metric.

Machine	Accuracy	Sensitivity	Specificity	Pos. Pred. Value	Neg. Pred. Value	MCC
MutationTaster	0.6502	0.8887	0.3198	0.6442	0.6748	0.2579
Polyphen	0.6543	0.8595	0.3699	0.6540	0.6553	0.2663
LRT	0.6607	0.7497	0.5376	0.6919	0.6078	0.2934
KVSNP*	0.8012	0.8886	0.4188	0.8699	0.4621	0.3194
SIFT	0.6801	0.8387	0.4605	0.6829	0.6732	0.3264
MutationAssessor	0.6842	0.7330	0.6166	0.7259	0.6250	0.3502
Provean	0.7124	0.8164	0.5684	0.7238	0.6909	0.3994
VEST3	0.7019	0.5633	0.8940	0.8804	0.5964	0.4669
3DMPP	0.8113	0.8260	0.7905	0.8408	0.7722	0.6196

* SCN5A mutations could not be analyzed by KvSNP.

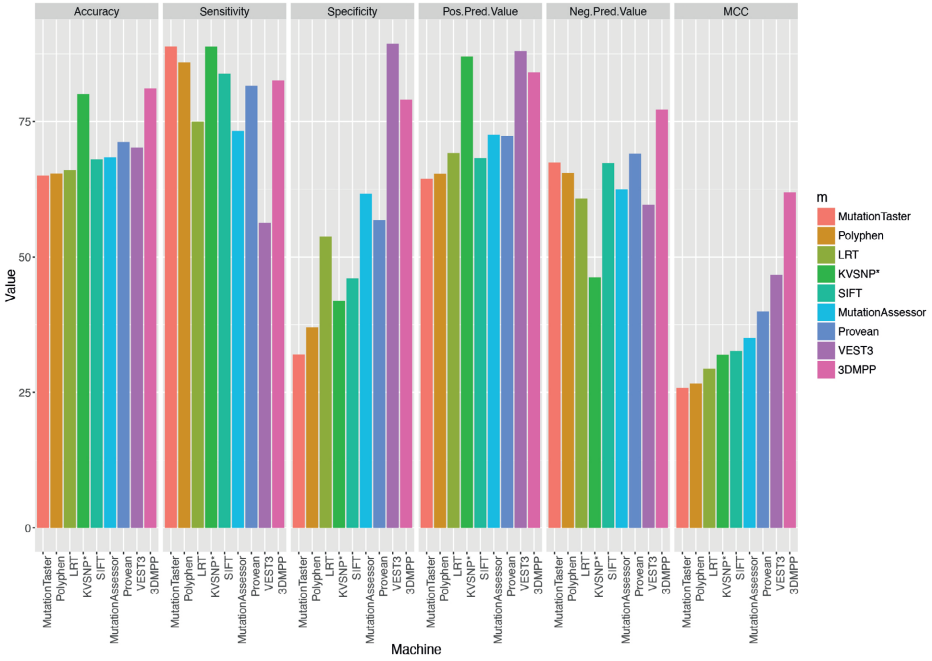


Figure 1. Comparison of the performance of the different predictors. KvSNP was unable to predict SCN5A. As such, results displayed here only reflect performance on the KCNH2 and KCNQ1 proteins.

The MCC, a good indicator of the overall performance of the individual tools[18], shows that the 3DMPP algorithm clearly outperforms all other predictors. As can be seen in table 3 and figure 1, especially in the negative predictive value, 3DMPP has the highest values compared to other predictors. All other prediction methods are often worse at

both the positive predictive value and the negative predictive value. Especially KVSNP and VEST3 appear biased towards assigning pathogenic indications to variants. The poor performance of KVSNP in predicting the effects of benign variants becomes clear when looking at the negative predictive values (NPV). This tool scores very close to 0.5, which indicates that the performance of these predictions is about the same, or slightly worse, than random. VEST3 uses HGMD presence as a predictive feature to assess known pathogenic mutants. This explains the relatively high scores obtained with the dataset used in this study since many of the pathogenic mutations are present in the HGMD. Therefore, on novel variants the positive predictive power of this tool will likely be lower than shown in this work. While the sensitivity of many predictors is often relatively high, the specificity is almost always much lower. This indicates that these predictors generated many false positive results.

4 DISCUSSION & CONCLUSION

Both Long QT Syndrome (LQTS) and Brugada syndrome (BrS) are potentially lethal disorders based on pathogenic variants in various ion channel genes. Clearly, for such sudden death–predisposing diseases for which highly effective medicine- (B-blockers), surgery- (left cardiac sympathetic denervation), and device-related therapies (ICD's, PM's) are available, accurate and proper interpretation of the genetic test is of paramount importance considering the diagnostic, prognostic, and therapeutic implications of a positive genetic test. Of equal importance is a potential negative test, including the identification of innocent variants, because individuals will be reassured and not further evaluated.

Unfortunately, linkage data, another indicator for pathogenicity, is rarely available. Functional studies are definitely able to support a causal relationship between the variant and the phenotype, but notable exceptions have been described[4]. Moreover, such studies are extremely time consuming, results in different heterologous cell models are quite often contradictory and most importantly, for only a great minority of identified variants functional data are available.

Molecular genetic data may also be of help. As a general concept a high prevalence of a given variant is indicative for non-pathogenicity. However, our previous research revealed particular locations within ion channel genes where pathogenic variants cannot be confidently differentiated from rare variants[27]. The estimated predictive value (the probability of a variant being pathogenic) could be as low as 0%, for example in large interdomain area's within SCN5a, the cardiac sodium channel gene.

The work presented here was developed in response to the need for high-quality predictions about the pathogenicity of nsSNPs in LQTS- and BrS-related genes. We show that 3DMPP is best able to separate disease-causing nsSNPs from benign nsSNPs

compared to other tools, which include both domain-specific predictors as well as genome wide predictors.

It has proven to be very difficult to create a dataset where it is beyond doubt that there are no pathogenic mutations in the benign dataset and vice versa. With a prevalence of 1:2000 for LQT and regular incomplete penetrance, it is more likely that our case and control populations are very enriched for pathogenic and benign variants, than that there is an absolute separation between the two sets. To what extend these inaccuracies in the reference sets have effects on the classification performance is unknown.

The accurate predictions made by 3DMPP shows that the availability of reliably annotated patient cohort data together with the use of high-quality, domain-specific integrated data is essential for creating a reliable prediction tool that performs well in assigning pathogenicity predictions to both benign and pathogenic nsSNPs.

The interpretation of variants flowing out of next-generation sequencing experiments is one of the largest bottlenecks for DNA diagnostics. As we have shown here, frequently used genome-wide prediction tools and even gene-family specific predictors are clearly not up to this task, thereby possibly misleading downstream experimental and clinical investigations. Consequently, there is an enormous need within the community for novel tools to reliably interpret the effects of polymorphisms for the complete genome.

To accommodate this, 3DM information systems are being created for the complete human proteome. All information systems contain the same types of information as the systems described above; extensive sequence and alignment-derived data, integrated with structural data and information extracted from literature. For all possible variants, pathogenicity predictions will be available.

The classifier described in this article has been trained and tested on ion channels. To what extend this classifier is applicable to other protein families and to what extend it is ion channel specific will need to be further investigated, but it is expected that newer versions that will be based on more and more diverse data will see increased performance. Although a daunting task, we believe this will be a major step forward for reliable routine diagnostics of unclassified variants on a large scale.

SUPPORTING INFORMATION

Supplementary data is available online:
https://figshare.com/s/0dc52d8b710ded5ceca3

AUTHOR CONTRIBUTIONS

Conceptualization: HJ BV TB. Data curation: SH BV TB RK. Formal analysis: SH BV TB. Funding acquisition: HJ AW MM. Investigation: BV SH TB. Methodology: SH BV TB MT MA. Project administration: HJ. Resources: HJ RD AW MM JK MA. Supervision: HJ RD AW MM. Validation: BV SH MA TB. Visualization: SH. Writing draft and editing: BV TB SH.

REFERENCES

1. Stenson, P. D. *et al.* The Human Gene Mutation Database: providing a comprehensive central mutation database for molecular diagnostics and personalized genomics. *Hum. Genomics* **4,** 69–72 (2009).
2. Cooper, D. N. *et al.* Genes, mutations, and human inherited disease at the dawn of the age of personalized genomics. *Hum. Mutat* **31,** 631–655 (2010).
3. Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452,** 872–876 (2008).
4. Watanabe, H. *et al.* Striking In vivo phenotype of a disease-associated human SCN5A mutation producing minimal changes in vitro. *Circulation* **124,** 1001–1011 (2011).
5. Ackerman, M. J. *et al.* HRS/EHRA Expert Consensus Statement on the State of Genetic Testing for the Channelopathies and Cardiomyopathies This document was developed as a partnership between the Heart Rhythm Society (HRS) and the European Heart Rhythm Association (EHRA). *Europace* **13,** 1077–1109 (2011).
6. Priori, S. G. *et al.* Executive summary: HRS/EHRA/APHRS expert consensus statement on the diagnosis and management of patients with inherited primary arrhythmia syndromes. *Europace* **15,** 1389–1406 (2013).
7. Kuipers, R. K. *et al.* 3DM: systematic analysis of heterogeneous superfamily data to discover protein functionalities. *Proteins* **78,** 2101–2113 (2010).
8. Song, W. *et al.* Exploring the landscape of pathogenic genetic variation in the ExAC population database: insights of relevance to variant classification. *Genet. Med.* (2015). doi:10.1038/gim.2015.180
9. Kuipers, R. K., Joosten, H.-J., Lekanne dit Deprez, R. H., Mannens, M. M. & Schaap, P. J. Novel tools for extraction and validation of disease-related mutations applied to Fabry disease. *Hum. Mutat* **31,** 1026–1032 (2010).
10. Consortium, T. 1000 G. P. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
11. Oliveira, L., Paiva, A. C. M. & Vriend, G. Correlated mutation analyses on very large sequence families. *Chembiochem* **3,** 1010–1017 (2002).
12. Oliveira, L., Paiva, P. B., Paiva, A. C. M. & Vriend, G. Identification of functionally conserved residues with the use of entropy-variability plots. *Proteins* **52,** 544–552 (2003).
13. Folkertsma, S. *et al.* A family-based approach reveals the function of residues in the nuclear receptor ligand-binding domain. *J. Mol. Biol* **341,** 321–335 (2004).
14. Kowarsch, A., Fuchs, A., Frishman, D. & Pagel, P. Correlated Mutations: A Hallmark of Phenotypic Amino Acid Substitutions. *PLoS Comput Biol* **6,** e1000923 (2010).
15. Kuhn, M. Building Predictive Models in *R* Using the **caret** Package. *Journal of Statistical Software* **28,** (2008).
16. Breiman, L. Random Forests. *Machine Learning* **45,** 5–32 (2001).
17. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **29,** 1189–1232 (200110).
18. Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. & Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **16,** 412–424 (2000).
19. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31,** 3812–3814 (2003).

20. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7,** 248–249 (2010).

21. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE* **7,** e46688 (2012).

22. Stead, L. F., Wood, I. C. & Westhead, D. R. KvSNP: accurately predicting the effect of genetic variants in voltage-gated potassium channels. *Bioinformatics* **27,** 2181–2186 (2011).

23. Carter, H., Douville, C., Stenson, P. D., Cooper, D. N. & Karchin, R. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* **14 Suppl 3,** S3 (2013).

24. Schwarz, J. M., Cooper, D. N., Schuelke, M. & Seelow, D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Meth* **11,** 361–362 (2014).

25. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucl. Acids Res.* gkr407 (2011). doi:10.1093/nar/gkr407

26. Chun, S. & Fay, J. C. Identification of deleterious mutations within three human genomes. *Genome Res* **19,** 1553–1561 (2009).

27. Kapa, S. *et al.* Genetic testing for long-QT syndrome: distinguishing pathogenic mutations from benign variants. *Circulation* **120,** 1752–1760 (2009).

CHAPTER 6

General discussion

1 SCOPE

As the amount of available protein data continues to increase exponentially, methods for the automated collection and integration are in high demand. Information systems that can store and interconnect these data in relation to the protein family provide an opportunity to view knowledge of individual proteins in perspective. We have used such integrated protein family data systems to analyze protein families in the context of human disorders and protein engineering. Furthermore, we developed applications for mutant pathogenicity prediction and to aid protein engineering. The collection and integration of data for complete protein superfamilies provides added value to data collections for individual proteins. This chapter discusses the research presented in this thesis and offers an outlook on future developments.

2 DATA COLLECTION AND STORAGE

With the development of new techniques in both the sequencing and protein structure elucidation fields it is likely that the rate at which protein data is generated will continue to increase. The collection and integration of such data is, therefore, an increasingly arduous task. This thesis mainly covers the integration of three different data sources; protein sequences, structures and literature. Compared to the literature, sequence and structural data are relatively straightforward to access and use. Sequence databases store huge amounts of sequence data that can be indexed and clustered with tools such as BLAST and HMMER to efficiently identify relevant sequences with sequence similarity searches (Altschul et al. 1990; Potter et al. 2018). Similarly, protein structures are stored and indexed in supplementary databases such as SCOP (Fox et al. 2014). Therefore, it can be presumed that the storage of these data can be scaled easily.

Beside the increase in number of protein structures solved each year, the quality of these structures continues to improve as well. This allows redundant low-quality structures to be replaced and structural alignments to be improved. Improvements in electron microscopy (EM) structure determination will allow for many new structures of both large and small proteins to be determined such as membrane proteins that have proven difficult to obtain (Vénien-Bryan et al. 2017; Shoemaker and Ando 2018). Additionally, these EM structures can provide protein-protein interaction data for protein complexes that could not be crystallized for X-ray structures. Such data, when added to protein family information systems, can help increase understanding of interaction sites and help relate structural conformation to biological function. Furthermore, when the number of available structures that contain substrates grows, insight could be gained into the function of existing substrates and thereby help the prediction of novel substrates for both biotech and drug design.

2.1 Text-mining

Compared to sequence and structural data, the collection and integration of biomedical literature is a more challenging task to perform automatically. To start, the identification of literature related to a protein family is precarious as keyword searches can result in either very few or millions of documents depending on field of study and keyword selection. Additionally, the retrieval of full-text papers is hampered by the fragmentation of literature resources and that many papers are not publicly accessible. Fortunately, the call for open access research has in recent years gained more attention (Sukhov et al. 2016). Funding agencies, such as the EU, increasingly require open access publications, often with the argument that research funded by the public domain should be accessible without restriction. In addition to advancements in automated text mining increased availability of research will also benefit researchers with less financial means.

The automatic extraction of knowledge from written text is a non-trivial task due to the variability of written text. In **chapter 2** a text-mining method is described that is limited to the extraction of mutation data from full-text papers. Mutations are typically described in easily recognizable formats yet reflect valuable experimental data. This simplifies the extraction process so that they can be identified and connected to proteins with relative ease. A method such as mutator shows a clear advantage over reading publications manually due to the number of relevant articles for any given protein family. **Chapter 2** describes the creation of a Fabry disease mutation database with Mutator that contained many more mutations than were present in manually annotated databases. The creators of one of these databases, the HGMD, responded in an editorial stating that while Mutator could identify a larger number of mutations from more articles, manual approaches were still more accurate (Stenson and Cooper 2010).

Of course, more than just mutation data can be extracted from scientific articles. Techniques such as named entity recognition (NER) combined with co-occurrence or natural language processing (NLP) methods enabled the development of tools that extract other data from articles such as protein-protein interactions, protein-compound interactions or disease phenotypes (Fleuren and Alkema 2015; Przybyła et al. 2016). Moreover, the scientific literature might not be the only literary source of protein knowledge that can be mined automatically. Patent documents can be a valuable resource as well, albeit knowledge in such documents is generally written in a more veiled manner. Whereas scientific publications are written with the incentive to showcase the accomplished work, patents often attempt to restrict the freedom to operate for other parties while the discovered protein knowledge they reveal is minimized. This offers patent specific challenges for text mining approaches. Nonetheless, the projection of extracted patent data on protein family alignments could potentially grant insight in patterns of disclosed information.

2.2 Data quality

We have arrived at a point where it can be questioned if more data is always better. In the past, every new sequence, structure or experiment added value and insight into a protein family. This is no longer necessarily true when each sequence added reveals less variation of a family's sequence space and thus provides less insight. UniProt already removed a large part of the sequences in their database in 2015 because they were derived from highly similar strains of bacterial species (Bateman et al. 2017).

More importantly, there have also been many examples of published protein data that afterwards were considered to be of bad quality. Partial sequences are often excluded from protein sequence databases. Structures with low resolutions can be ignored as higher quality versions become available. Likewise, structures that were released with elucidation mistakes can be corrected. The PDB-REDO databank contains such optimized crystallographic structures (Joosten et al. 2014). Human variants discovered in association studies and described in the literature as pathogenic turned out to be benign natural variants (Niemann et al. 2013). Therefore, increased quality control of data is needed so that ultimately the quality of systems that integrate and make use of this data is improved. While the generation of more protein related data will undoubtedly grant new insights into protein function, simultaneously it allows for an increasingly selective strategy to collect and integrate higher quality data.

3 INTEGRATED DATA ANALYSIS

The storage of protein family data in itself only provides easy access for manual inspection. It is the interconnection of these data between protein family members that allows observations to be placed in the context of the protein family to gain additional knowledge and insight into protein function. Protein information systems such as 3DM connect equivalent protein residues across protein family members to combine and compare residue annotations. This allows novel insights to be gained that are hard or impossible to derive merely by manual inspection. By the utilization of data from protein family members, experimental efforts can be directed more precisely and efficiently.

In **chapter 2** we observed that the gathered pathogenic mutations leading to Fabry's disease are biased toward regions of the protein that can be allocated to the structurally conserved core and toward residues that are not abundant in nature. Additionally, in protein engineering experiments it was observed that the introduction of residues that do not occur in superfamily alignments can be limited to minimize mutagenesis library sizes (Jochens and Bornscheuer 2010; Jochens et al. 2010).

In a further investigation of this protein family in **chapter 3,** we observed that the positions with pathogenic mutations overlap with homologs even over very large phylo-

genetic distances. This indicates that known variants of remote homologs can play a role in the prediction of (pathogenic) effects for novel variations.

Chapter 4 describes the application of correlated mutation analysis to identify networks of alignment positions that co-evolved throughout evolution to preserve certain functions. In this chapter an effort is made to reveal the function of such networks through the combination of network positions with knowledge automatically mined from the literature. Additionally, we showed that assigned functions can be transferred to previously undescribed positions in such a network.

Chapters 2-4 show that protein family information systems can provide new insights that can result only from the combination of data derived from many individual proteins.

Many of these analyses simply generate new data derived from integrated data, however, they do not necessarily provide direct insight into the biology of proteins without additional manual investigation and deduction. Superfamily conservation for example, can reveal that a residue is important yet not why this residue is important. Correlated mutation analysis (CMA) of an alignment results only in a set of co-evolving positions, and these positions are heavily dependent on the phylogenetic scope of the input alignment. The resulting networks can still be considered as abstract data as they do not reveal the biological functions that they're responsible for. Even keyword-related positions can still be considered as abstract data since keywords can be relatively abstract as well. It is only when these data are combined that a biological function can sometimes be revealed. Otherwise a manual inspection and deduction step is still required to effectively use this data as a clue to gain insight into protein function.

What other opportunities exist where we can combine data derived from integrated protein information systems to help understand the biology of proteins? A subsequent development could be to apply CMA across multiple protein families to reveal protein-protein interaction interfaces and pathways. Such an analysis does require high-quality annotations of protein sequences that can be paired across different alignments.

4 APPLICATIONS OF INTEGRATED PROTEIN DATA SYSTEMS

Chapters 4 and **5** show that practical applications can be designed on top of protein family information systems such as 3DM. A small but smart mutation library to change enantioselectivity was manually designed as described in **chapter 4.** However, the steps to design proteins could be semi-automated in a tool that builds on top of 3DM. First, the selection of potential target proteins could be aided by a tool that makes use of the superfamily alignment to take sequence motifs, phylogenetic characteristics and other protein properties into account. Second, the design of mutants could benefit from both

the alignment and alignment derived data such a co-evolution networks and literature keyword related positions.

Nobili et al. applied the CorNet tool described in **chapter 4** to engineer an esterase from *Pseudomonas fluorescence* (PSE). They identified hotspots and potential mutant residues with the use of a 3DM system for the a/b-hydrolase fold to develop a small mutant library. This library of <80 mutations resulted in a mutant with 15-fold improved enantioselectivity (Nobili et al. 2015). Genz et al. engineered the substrate specificity of a class I pyridoxal-51 -phosphate (PLP)-dependent enzyme. With a 3DM system for this superfamily, positions related to specificity were identified and targeted with a small set of hydrophobic residues. This library of 1,600 variants resulted in three hits that could perform an aldehyde transferase reaction (Genz et al. 2015). Junker et al. engineered an aldolase with a complete switch of reaction specificity with a small library that focused on three positions, two of which showed correlated mutation behavior (Junker et al. 2018). In another study, Knight et al. used the 3DM system for class III PLP-dependent enzymes to discover a new multimeric racemase by expressing four uncharacterized proteins that had a low sequence similarity (but high structural similarity) with known alanine racemase and decarboxylases (Knight et al. 2017).

Alternatively, **chapter 5** shows the development of a human variant pathogenicity predictor that, based on better integrative data, can outperform competitors and aid in the diagnosis and treatment of diseases. Such machine learning models make use of patterns from data available for known variants to predict how novel variations affect the phenotype. While this field of artificial intelligence advances rapidly and provides useful applications, the process of decisions made by these applications can often still be considered as a black box. The understanding and description of what causes a variant to be pathogenic in the real world is lacking from these predictors. Even when a clear explanation can be given of which patterns in the data determined the outcome, the biology behind it is not explained. A better notion of how variations lead to disease is essential to eventually understand the biology.

In Fabry's disease, discussed in **chapters 2** and **3**, mutations lead to a defective enzyme that causes the buildup of a glycolipid that impairs various tissue functions. In other cases, mutations in the binding interface of a receptor protein can disrupt signal pathway transduction and gene transcription (McDonald et al. 2000). However, for a predictor to distinguish a pathogenic mutation from a mutation that is detrimental to just a single protein, information on mechanisms that can (partly) replace a protein's function is required (MacArthur et al. 2012). Currently, the explanation of the biology behind pathogenic variants is still performed by hand, this is hampered as the predictions are produced by a black box. A step toward better explanations of the biology is to supplement predictions with expert knowledge, as attempted by HOPE (Venselaar et al. 2010). To enable machine learned models to better explain decisions, better annota-

tions of protein residues and high-quality data will be required. Simultaneously, when high-quality annotations and patterns extracted from integrated protein family data become more readily available, ligand binding interfaces for example could be better understood. This would enable the development of better tools for the engineering of proteins with modified substrates or the design of inhibiting drugs.

5 FUTURE PERSPECTIVES

We have entered the age of large-scale whole genome sequencing (1000 Genomes Project Consortium et al. 2015). Soon, we will have the ability to sequence every (newborn) human to detect and analyze variations. Eventually, this will enable early detection of genetic disorders so that treatment can potentially start earlier to minimize harm inflicted by disease. However, this will undoubtedly lead to the discovery of many novel genetic variations (Shendure et al. 2017). One of the main subsequent problems will likely be to distinguish irrelevant natural diversity from pathogenic mutations. Therefore, high quality genome wide pathogenicity predictors will be necessary to handle this forecasted increase of encountered human genetic variations. Currently, the treatment of symptoms is often the only option when diseases are discovered. New technologies such as the Crispr-Cas system are currently in development that will potentially allow us to repair unwanted genetic variations (Razzouk 2018). However, it will still take many years of research to ensure these techniques are precise enough before human applications can be considered (Kosicki et al. 2018). Nevertheless, before such genetic editing tools have matured enough for human applications the technology to adequately assess genetic variations should be available.

The pathogenicity predictor described in **Chapter 5** that outperforms competitors, is available only for a limited set of proteins that belong to a single protein class. To support the development of whole genome high quality predictors, protein information systems targeted on all human proteins should be developed to enable machine learning techniques to optimally train predictors for novel genetic variants.

Similar challenges are encountered in the biotechnology field. This industry is growing rapidly, supported by an increase in discovered enzymes (Yang and Ding 2014). The engineering of enzymes for the degradation of plastics (Austin et al. 2018), or the production of hydrogen for green energy (Kubas et al. 2017) are just two examples of enzymes that represent currently relevant opportunities.

While the increasing amount of sampled sequences likely contain many interesting proteins for biotech applications, the data needs to be managed properly so that potentially interesting proteins can be identified among the large amounts of less suitable proteins.

To support the discovery and development of enzyme applications, protein engineering systems should offer new tools to select and optimize proteins. Such tools can benefit from protein family systems to optimally make use of patterns derived from integrated superfamily alignments. For example, machine learned models that help predict the fitness of proposed variants could help to more efficiently engineer a new generation of enzymes.

Scaling up 3DM

Both the diagnostics and protein engineering fields can benefit from the availability of protein family information systems to better understand proteins. Proteins targeted for investigation cover an increasing number of protein families and thus require an expansion of such systems available today. New higher-quality protein data become available every year. Applications build on integrative protein information systems would benefit from the integration in this data. Furthermore, as the number of targeted proteins increases and new data becomes available at a higher rate, the generation of protein family information systems will have to be automated so that the integrated data available is kept up to date. Eventually, automatically generated protein information systems for all protein families will be required. Therefore, there is a lot of work ahead to make the large-scale generation of protein family information systems possible. To conclude, it remains an ongoing effort to manage the quantity and quality of generated protein data to help understand the relationship between protein sequence, structure and function.

REFERENCES

1000 Genomes Project Consortium, Auton A, Brooks LD, et al (2015) A global reference for human genetic variation. Nature 526:68-74. doi: 10.1038/nature15393

Altschul SF, Gish W, Miller W, et al (1990) Basic local alignment search tool. J Mol Biol 215:403-410. doi: 10.1016/S0022-2836(05)80360-2

Austin HP, Allen MD, Donohoe BS, et al (2018) Characterization and engineering of a plastic-degrading aromatic polyesterase. PNAS 201718804. doi: 10.1073/pnas.1718804115

Bateman A, Martin MJ, O'Donovan C, et al (2017) UniProt: the universal protein knowledgebase. Nucleic Acids Res 45:D158-D169. doi: 10.1093/nar/gkw1099

Fleuren WWM, Alkema W (2015) Application of text mining in the biomedical domain. Methods 74:97-106. doi: 10.1016/j.ymeth.2015.01.015

Fox NK, Brenner SE, Chandonia J-M (2014) SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res 42:D304-D309. doi: 10.1093/nar/gkt1240

Genz M, Vickers C, Van den Bergh T, et al (2015) Alteration of the Donor/Acceptor Spectrum of the (S)-Amine Transaminase from Vibrio fluvialis. International Journal of Molecular Sciences 16:26953-26963. doi: 10.3390/ijms161126007

Jochens H, Aerts D, Bornscheuer UT (2010) Thermostabilization of an esterase by alignment-guided focussed directed evolution. Protein Eng Des Sel 23:903-909. doi: 10.1093/protein/gzq071

Jochens H, Bornscheuer UT (2010) Natural diversity to guide focused directed evolution. ChemBioChem 11:1861-1866. doi: 10.1002/cbic.201000284

Joosten RP, Long F, Murshudov GN, Perrakis A (2014) The PDB_REDO server for macromolecular structure model optimization. IUCrJ 1:213-220. doi: 10.1107/S2052252514009324

Junker S, Roldan R, Joosten H, et al (2018) Complete Switch of Reaction Specificity of an Aldolase by Directed Evolution In Vitro: Synthesis of Generic Aliphatic Aldol Products. Angew Chem Int Ed Engl 57:10153-10157. doi: 10.1002/anie.201804831

Knight AM, Nobili A, van den Bergh T, et al (2017) Bioinformatic analysis of fold-type III PLP-dependent enzymes discovers multimeric racemases. Appl Microbiol Biotechnol 101:1499-1507. doi: 10.1007/s00253-016-7940-7

Kosicki M, Tomberg K, Bradley A (2018) Repair of double-strand breaks induced by CRISPR-Cas9 leads to large deletions and complex rearrangements. Nature Biotechnology. doi: 10.1038/nbt.4192

Kubas A, Orain C, Sancho DD, et al (2017) Mechanism of O_2 diffusion and reduction in FeFe hydrogenases. Nature Chemistry 9:88-95. doi: 10.1038/nchem.2592

MacArthur DG, Balasubramanian S, Frankish A, et al (2012) A systematic survey of loss-of-function variants in human protein-coding genes. Science 335:823-828. doi: 10.1126/science.1215040

McDonald S, Brive L, Agus DB, et al (2000) Ligand responsiveness in human prostate cancer: structural analysis of mutant androgen receptors from LNCaP and CWR22 tumors. Cancer Res 60:2317-2322

Niemann M, Rolfs A, Giese A, et al (2013) Lyso-Gb3 Indicates that the Alpha-Galactosidase A Mutation D313Y is not Clinically Relevant for Fabry Disease. JIMD Rep 7:99-102. doi: 10.1007/8904_2012_154

Nobili A, Tao Y, Pavlidis IV, et al (2015) Simultaneous use of in silico design and a correlated mutation network as a tool to efficiently guide enzyme engineering. Chembiochem 16:805-810. doi: 10.1002/cbic.201402665

Potter SC, Luciani A, Eddy SR, et al (2018) HMMER web server: 2018 update. Nucleic Acids Res 46:W200–W204. doi: 10.1093/nar/gky448

Przybyła P, Shardlow M, Aubin S, et al (2016) Text mining resources for the life sciences. Database (Oxford) 2016:. doi: 10.1093/database/baw145

Razzouk S (2018) CRISPR-Cas9: A cornerstone for the evolution of precision medicine. Ann Hum Genet. doi: 10.1111/ahg.12271

Shendure J, Balasubramanian S, Church GM, et al (2017) DNA sequencing at 40: past, present and future. Nature 550:345-353. doi: 10.1038/nature24286

Shoemaker SC, Ando N (2018) X-rays in the Cryo-EM Era: Structural Biology's Dynamic Future. Biochemistry 57:277-285. doi: 10.1021/acs.biochem.7b01031

Stenson PD, Cooper DN (2010) Prospects for the automated extraction of mutation data from the scientific literature. Hum Genomics 5:1-4

Sukhov A, Burrall B, Maverakis E (2016) The history of open access medical publishing: a comprehensive review. Dermatol Online J 22:

Vénien-Bryan C, Li Z, Vuillard L, Boutin JA (2017) Cryo-electron microscopy and X-ray crystallography: complementary approaches to structural biology and drug discovery. Acta Crystallogr F Struct Biol Commun 73:174-183. doi: 10.1107/S2053230X17003740

Venselaar H, Te Beek TAH, Kuipers RKP, et al (2010) Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. BMC Bioinformatics 11:548. doi: 10.1186/1471-2105-11-548

Yang G, Ding Y (2014) Recent advances in biocatalyst discovery, development and applications. Bioorg Med Chem 22:5604-5612. doi: 10.1016/j.bmc.2014.06.033

6

SUMMARY

Proteins are key to many different processes in and around living cells, they are the molecular mechanisms that facilitate life. Therefore, protein function and structure have been of great interest to researchers over the past decades. A better understanding of how proteins work is important, as it will help us understand the various diseases caused by their malfunction. Yet they can also be used in various applications, for example as drug targets or as drugs themselves to treat disease. Proteins have other, non-medical applications, they can for example be utilized as catalysts in laundry detergent or industrial processes. The research on proteins has resulted in an ever-increasing amount of data related to proteins.

Protein information systems were developed to make sense out of all the data available for specific protein classes. The work described in this thesis makes use of the 3DM platform to store and order the enormous amount of protein data that is nowadays available. Therefore, in the introduction (**chapter 1**) a technical description of the generation of 3DM systems is provided. The 3DM platform integrates heterogeneous datasets for complete protein superfamilies to generate molecular class specific information systems (3DM systems). A structure based multiple sequence alignment connects proteins that share a similar 3-dimensional structure or sequence, and a unifying residue numbering scheme allows for protein data to be combined and investigated for all homologous proteins present in a 3DM system.

The main objective of my PhD thesis is to facilitate a better understanding of protein function through data integration- and analysis-tools to support both DNA diagnostics and protein engineering efforts. To that end, **chapter 2** and **chapter 3** focus on the collection and integration of mutational data from scientific literature, followed by protein superfamily analysis to reveal predictive patterns. In **chapter 4**, mutational data mined for whole families are combined with a superfamily alignment analysis to assign function to networks of residues that appear to have co-evolved. And in **chapter 5**, the data of a 3DM system is used to train a machine learning model for the prediction of pathogenicity of novel genetic variants.

In **chapter 2**, a method for the automated collection and integration of mutation data from scientific literature is described. This text-mining tool, Mutator, was designed to automatically identify and retrieve relevant literature for a complete protein family. Mutations are extracted with pattern recognition, assigned to proteins through a grounding process and consecutively integrated in the 3DM system.

Mutator was tested on the alpha-amylase protein family which contains a protein that, when mutated, can lead to Fabry's disease, a monogenetic lysosomal storage disorder.

The Fabry disease mutation database was constructed which contained 30% more mutation data when compared to the manually curated HGMD.

To differentiate between natural variation and pathogenic mutations, alignment patterns in the alpha-amylase superfamily were investigated and Validator was developed, a tool to evaluate the potential pathogenicity of individual mutations using statistics such as conservation and mutant side chain accessibility.

In a follow up study, **chapter 3**, nine disease-related proteins across two different protein families are investigated. An improved version of Mutator was developed and used to collect mutational data for these protein families. A comparison to the manually curated HGMD showed that Mutator extracts far more mutation data, both in terms of unique mutations extracted, as well as number of references per mutation.

These mutation datasets were analyzed in the context of the protein families to identify patterns to assess pathogenicity for novel variants. This analysis showed that pathogenic mutations would more often be observed in the structurally conserved core of the family than in structurally variable regions. I also observed a significant overlap of protein positions with pathogenic mutations between very distant homologs, indicating that data can be transferred between equivalent residue positions even across large phylogenetic distances. The sequence identity between the compared proteins is as low as 10% and thus these proteins can no longer be aligned by conventional sequence alignment programs. Protein superfamily information systems that use structural alignments, such as 3DM, can facilitate the transfer of data across equivalent residues of structurally conserved positions. Therefore, the availability of 3DM systems can be a valuable asset for predictive models, as it can provide predictive features that otherwise would be unattainable, such as mutation data for very distant homologs.

Chapter 4 describes a web-based tool for the analysis of co-evolving residues in protein super-family alignments. This tool, CorNet, combines networks of co-evolving positions with literature data to shed light on the function of these residues. When keywords, such as “specificity”, are mentioned in the sentences that contain mutational data for an alignment position, these were used to annotate the alignment position of a network.

For six different protein families, a list of literature keywords was used to annotate network positions and to calculate an enrichment score of any keyword to be specifically associated with a network of co-evolving positions. Enriched keywords indicate the function behind evolutionary pressures that caused networks of residues to co-evolve. It was shown that an assigned function can be transferred from annotated to non-annotated positions in a network with a mutagenesis study that targeted only the non-annotated positions of a network enriched for the keyword “enantioselectivity”. Furthermore, changing the scope of proteins that make up the alignment results in

different networks of co-evolving positions, this provides an opportunity to finetune the alignment of a 3DM system to identify residue networks that perform a specific function.

In **chapter 5**, we utilize the integrated data of 3DM systems to train a model to predict variant pathogenicity with machine learning techniques. For three genes related to Long QT syndrome and Brugada syndrome, both cardiac arrhythmia syndromes, datasets of both benign and pathogenic variants were collected. 3DM was used to generate much heterogeneous data to describe each variant. Using this data, a pathogenicity predictor (3DMPP) was generated to separate disease-causing from non-disease causing variants. 3DMPP outperformed frequently used genome-wide prediction tools and even gene-family specific predictors. However, in this study it proved to be very difficult to create a high-quality training dataset as different sources can disagree on variant annotation. Especially considering therapeutic implications of a genetic test, the 3DMPP results shows that a reliably annotated patient cohort dataset combined with domain-specific structure based integrated data is essential for creating a reliable prediction tool.

Finally, in the general discussion (**chapter 6**), I address the broader context and discuss the future perspectives of the tools described in this thesis. I discuss the access, retrieval, quality, and storage of the continuously increasing amounts of protein data such as the shortcomings and opportunities in text mining and the push towards open access scientific literature. I reflect on the applications described in the research chapters of this thesis and discuss new applications of integrated protein family information systems and several examples are given of engineering projects that made use of 3DM systems in various ways, including the tools developed in this thesis. And finally, an outlook is given on the scale-up of protein information systems to cover the complete structural space and even the whole human exome.

DANKWOORD

Het heeft even geduurd, maar we zijn er. Terugkijkend is het ook zeer terecht om hier de mensen te bedanken die mij de afgelopen jaren gesteund en geholpen hebben. Nu het uiteindelijk toch gelukt is ben ik erg dankbaar voor alle hulp die ik heb gehad van vrienden, collega's en mijn promotoren.

Vitor, uiteraard wil ik jou als promotor bedanken voor het mogelijk maken van mijn promotie. Het is bijzonder om zo'n industrieel PhD traject te doorlopen – uiteraard met voor- en nadelen – maar uiteindelijk is het toch gelukt om dit succesvol af te ronden. Gert, mijn tweede promotor, ik kan mij de dagen schrijvend aan jouw keukentafel nog goed herinneren en ben erg dankbaar voor je hulp over de afgelopen jaren. Ik hoop dat je erg kan genieten van de Filipijnen en af en toe nog wat bugs in WHATIF wilt blijven fixen voor ons ;)

Ook het team bij Bio-Prodict ben ik erg dankbaar. Dit avontuur waar ik 15 jaar terug tijdens mijn bachelor stage al bij betrokken werd heeft geen moment verveeld. Eerst met z'n drieën in de kelder van de WUR en nu met ruim 15 man in ons kantoor in Nijmegen. Henk-Jan, jij verdient een prominente plek in mijn dankwoord. Door het mogelijk maken van onze samenwerking met de WUR heb ik uiteindelijk dit boekje kunnen schrijven.

Bas, jou ben ik ook erg dankbaar voor je hulp tijdens mijn promotietraject, maar ook zeker voor je inzichten en de vele gesprekken over de politiek, de ECB en de huizenmarkt die we de afgelopen jaren hebben gevoerd. En uiteraard moet ik ons Nijmeegse tafelvoetbal kampioenschap even noemen, hopelijk kunnen we snel onze titel verdedigen.

Remko, het beachvolleybal na werk blijft een geweldige hobby, het is heerlijk om in het zand al het dagelijks gedoe even van ons af te zetten. Heel fijn ook dat ik regelmatig bij jou heb kunnen overnachten na een late training nu ik niet meer in Nijmegen woon.

Ook mijn andere collega's Hidde, Olga, Jochem, Sergio, Daniël, Julian, Stephan en natuurlijk onze recente aanwinsten Jeanine en Maarten; jullie zijn geweldig. De relaxte sfeer, de toffe uitjes en het fanatieke tafelvoetbal maken Bio-Prodict wat is. Joanna, although you have now moved on to a new position closer to home, I really treasured the way we worked together, and I hope we can one day resume that.
I also want to include Thomas here, you're the best "BD as a service" guy I know ;)

Wat onze werkuitjes betreft; of het nu wakeboarden, parachutespringen, wintersport of paintball is, of vliegen met een stuntpiloot, ik geniet er vol van en hoop deze uitjes nog vaak met jullie te kunnen beleven!

Gelukkig heb ik de afgelopen jaren naast werk en het schrijven van dit boekje ook erg veel lol beleefd met oude vrienden en studiegenoten uit Zoetermeer, Leiden, Wageningen en Nijmegen.

Sander, wij hebben elkaar leren kennen op de studie Bio-informatica in Leiden. De feestjes en vooral de Koninginnedagen in die tijd zal ik niet gauw vergeten. Ik vind het altijd weer leuk om je te zien en te horen waar je mee bezig bent.

Wytze, jou ken ik toch wel het langste van iedereen, sinds we van kleins af aan tegenover elkaar woonden en we elk jaar met oud en nieuw wekenlang helemaal gek waren van vuurwerk. Ik geniet nog steeds erg van onze jaarlijkse etentjes, ik hoop dat we die nog lang kunnen vol houden!

Stefan, vroeger gingen we altijd voetballen na school en jij was vaak net iets beter. Daarna als jonkies samen naar Rome, Barcelona en Madrid de steden van Europa verkennen. Nu we allebei vader zijn geworden en een huis hebben gekocht is het leven toch wel anders. Hopelijk kunnen we straks met de kleine jongens op pad!

Nick, ten eerste bedankt voor de hulp met de cover van dit boekje. Wij kennen elkaar al sinds we op de basisschool bij jou thuis vaak Nintendo speelden. Ook onze (tweede) road trip zal ik ook nooit vergeten, waar we op dag één in Frankrijk al een gigantische pion onder de auto vandaan konden vissen omdat de remmen het niet meer deden. Toen ik ook in Utrecht kwam wonen konden we weer makkelijker afspreken en van de etentjes met jou en Lisette heb ik altijd erg genoten. Dat we die nog vaak kunnen herhalen!

Mart en Susan, mijn oude huisgenootjes uit Renkum. De gezellige tijd daar maakte mijn studententijd helemaal af en ik ben ontzettend blij dat we elkaar nog steeds zien, kom maar op met die wintersport!
Susan, sinds we de afgelopen jaren allebei in Utrecht woonden zijn we nog vaak gaan hardlopen. Ook nu ik naar Ede verhuis blijf je altijd welkom om samen door het bos te rennen. Mart, de hiking trips naar Ierland, Duitsland, Engeland en Schotland waren toch wel de hoogtepunten van mijn tijd in Wageningen. En oh ja, nogmaals sorry dat ik over je heen ben gefietst, het was gewoon erg donker tussen Wageningen en Renkum.

Dan de vriendengroep uit Nijmegen; Martin, Thomas, Lex, Tim, Joep, Eugene en Rob. Onze spelletjesavonden mogen dan vaak uitlopen op verhitte discussies, ik geniet er elke keer van! Ook de jaarlijkse tripjes naar Europa's beste C-locaties blijven legendarisch, ik hoop er nog veel mee te maken. Thomas, dat je nog vaak onze spelletjes avonden interessant mag maken. Tim, hopelijk mag je snel weer mistery trips voor ons organiseren! Martin, aka Tims chauffeur, alleen met jou durf ik mysterieuze deuren op festivals te openen ;)

Rob, onze YouTube ster, hopelijk kunnen we snel op bezoek komen in Wenen om eens te zien wat je daar allemaal stiekem uitspookt.
Joep, sinds je naar Nieuw-Zeeland verhuisd bent zien we je vooral digitaal, hopelijk kunnen we je snel weer eens in het echt zien. Onze StarCraft II avondjes samen met Eugene hebben de pandemie isolatie een stuk aangenamer gemaakt.

Lex, sinds de master in Wageningen kennen wij elkaar en over de afgelopen jaren hebben we van alles beleefd. Ik geniet er ook altijd weer van om met jou, Emily, Sophia en de kids af te spreken en te zien hoe het met jullie gaat.

Mijn paranimfen, Mart en Lex, jullie hebben mijn studietijd in Wageningen toch wel onvergetelijk gemaakt en daarmee hebben jullie terecht de positie als paranimf verdient. We moeten nog zien hoe het gaat met de lockdown, maar ik hoop toch dat we samen in rokkostuum naar de verdediging kunnen!

Ook alle familie wil ik graag bedanken voor de fijne tijd samen de afgelopen jaren.
De weekendjes weg, feestdagen, verjaardagen en spelletjesavonden, altijd fijn om langs te komen. Mam, pap, Aty, Marian en Gijs, ik vind het geweldig om te zien hoe jullie een band met Oskar opbouwen en jullie zijn een grote steun. Zusje (en Rob), hopelijk kunnen we snel weer op bezoek komen in Canada!

Sophia, je hebt mij altijd gesteund om door te zetten met dit promotie traject. Ik heb ook enorm genoten van de mooie reizen die we hebben gemaakt en het leven dat we samen hebben opgebouwd. Met de komst van Oskar is dat behoorlijk veranderd en heel wat hectischer geworden. De deadline om voor zijn geboorte een boekje te hebben is niet helemaal gehaald, maar dat geeft niet. Het is alleen *iets* lastiger thuiswerken. Maar hij maakt ons zo blij en het blijft bijzonder om te zien hoe hij zich ontwikkelt. Het blijft mijn favoriete moment van de dag als we 's avonds samen bij zijn bedje even kijken hoe hij ligt te slapen. Nu zijn we ook nog eens druk met ons nieuwe huis in Ede, ik kijk er ontzettend naar uit! Zonder jou had ik het niet gekund.

LIST OF PUBLICATIONS

Kuipers R*, **van den Bergh T***, Joosten H-J, Lekanne dit Deprez RH, Mannens MM, Schaap PJ. Novel tools for extraction and validation of disease-related mutations applied to Fabry disease. Hum Mutat. 2010 Sep;31(9):1026–32.

Joosten H-J, Kuipers RKP, **van den Bergh T**, Vriend G, Schaap PJ, Smit M. 3DM Protein Engineering Super-Family systems applied to the P450 family. In: 17th International Conference on Cytochrome P450: Biochemistry, Biophysics and Structure. Manchester; 2011.

Seddon G, Lounnas V, McGuire R, **van den Bergh T**, Bywater RP, Oliveira L, Vriend G. Drug design for ever, from hype to hope. J Comput Aided Mol Des. 2012 Jan;26(1):137–50.

Nobili A, Tao Y, Pavlidis IV, **van den Bergh T**, Joosten H-J, Tan T, Bornscheuer UT. Simultaneous use of in silico design and a correlated mutation network as a tool to efficiently guide enzyme engineering. Chembiochem. 2015 Mar 23;16(5):805–10.

Steffen-Munsberg F, Vickers C, Kohls H, Land H, Mallin H, Nobili A, Skalden L, **van den Bergh T**, Joosten H-J, Berglund P, Höhne M, Bornscheuer UT. Bioinformatic analysis of a PLP-dependent enzyme superfamily suitable for biocatalytic applications. Biotechnol Adv. 2015 Oct;33(5):566–604.

Genz M, Vickers C, **Van den Bergh T**, Joosten H-J, Dörr M, Höhne M, Bornscheuer UT. Alteration of the Donor/Acceptor Spectrum of the (S)-Amine Transaminase from Vibrio fluvialis. International Journal of Molecular Sciences. 2015 Nov;16(11):26953–63.

Genz M, Melse O, Schmidt S, Vickers C, Dörr M, **van den Bergh T**, Joosten H-J, Bornscheuer UT. Engineering the Amine Transaminase from Vibrio fluvialis towards Branched-Chain Substrates. ChemCatChem. 2016;8(20):3199–202.

Van Overtveldt S, Verhaeghe T, Joosten H-J, **van den Bergh T**, Beerens K, Desmet T. A structural classification of carbohydrate epimerases: From mechanistic insights to practical applications. Biotechnol Adv. 2015 Dec;33(8):1814–28.

Bergh T van den, Vroling B, Kuipers RKP, Joosten H-J, Vriend G. Common Pitfalls and Novel Opportunities for Predicting Variant Pathogenicity. Biochem Physiol. 2016 Feb 3;5(197).

Beier A, Bordewick S, Genz M, Schmidt S, **van den Bergh T**, Peters C, Joosten H-J, Bornscheuer UT. Switch in Cofactor Specificity of a Baeyer-Villiger Monooxygenase. Chembiochem. 2016 Dec 14;17(24):2312–5.

van den Bergh T*, Tamo G*, Nobili A, Tao Y, Tan T, Bornscheuer UT, Kuipers RKP, Vroling B, de Jong RM, Subramanian K, Schaap PJ, Desmet T, Nidetzky B, Vriend G, Joosten H-J. CorNet: Assigning function to networks of co-evolving residues by automated literature mining. PLoS ONE. 2017;12(5):e0176427.

Knight AM, Nobili A, **van den Bergh T**, Genz M, Joosten H-J, Albrecht D, Riedel K, Pavlidis IV, Bornscheuer UT. Bioinformatic analysis of fold-type III PLP-dependent enzymes discovers multimeric racemases. Appl Microbiol Biotechnol. 2017 Feb;101(4):1499–507.

Lanfranchi E, Pavkov-Keller T, Koehler E-M, Diepold M, Steiner K, Darnhofer B, Hartler J, **Van Den Bergh T**, Joosten H-J, Gruber-Khadjawi M, Thallinger GG, Birner-Gruenberger R, Gruber K, Winkler M, Glieder A. Enzyme discovery beyond homology: a unique hydroxynitrile lyase in the Bet v1 superfamily. Sci Rep. 2017 May 3;7:46738.

* shared first authorship

OVERVIEW OF COMPLETED TRAINING ACTIVITIES

Discipline specific activities

The molecular life of diatoms, Paris, France	2013
International Masterclass: Protein Engineering of Cytochrome P450s, Groningen	2013
Industrial Expert Workshop "Harvesting Environmental Genomes for the Development of Biocatalysts", Groningen	2013
Netherlands Bioinformatics Conference, Lunteren	2014
NBC-15 "Biotechnology by Dutch Design", Ede	2014
BioSB 2015, Lunteren	2015
Rsg retreat at biosb 2015, Lunteren	2015
16th European Congress on Biotechnology, Edinburgh, United Kingdom	2014
4th International Conference on Novel Enzymes, Ghent, Belgium	2014
Industry Expert Workshop: Marine Micr'Omics for Biotech Applications, Madrid, Spain	2015
International Masterclass: Computational Approaches for Discovery and Engineering of Enzymes for Biocatalysis and Synthetic Biology, Groningen	2015
European Conference on Marine Natural Products, Glasgow, United Kingdom	2015
pre-conference workshop of the ECMNP, Glasgow, United Kingdom	2015
B-wise Seminar, Wageningen	2015
Benelux Bioinformatics Conference, Nijmegen	2012
Horizon 2020 EU project Carbazymes 1st 3DM course, Nijmegen	2015
Horizon 2020 EU project Carbazymes 2nd 3DM course, Nijmegen	2015

General courses

Pattern Recognition, NBIC PhD school	2011
Computing for Data Analysis, Johns Hopkins University (Coursera)	2014
The Analytics Edge, MIT (edX)	2014

Other activities

Preparation of research proposal	
FP 7 EU project Kyrobio 3DM course, Graz, Austria	2012
FP 7 EU project MicroB3 3DM course, Nijmegen	2013
FP 7 EU project SuSy 3DM course, Nijmegen	2013
FP 7 EU project BioOx 3DM course, Nijmegen	2014
FP 7 EU project SuSy dissemination 3DM course, Nijmegen	2016

The research described in this thesis was financially supported by Bio-Prodict BV.

Printing: Gildeprint Enschede, gildeprint.nl
Layout and design: Anna Bleeker, persoonlijkproefschrift.nl
Cover design: Nick Poldermans