From existing data to novel hypotheses

Design and application of structure-based

Molecular Class Specific Information Systems

Remko Kuipers

Thesis committee

Thesis supervisors
Prof. dr. ir. V.A.P. Martins dos Santos
Professor of Systems and Synthetic Biology
Wageningen University

Prof. dr. G. Vriend
Professor of Bioinformatics of Macromolecular Structures
CMBI, Radboud University Nijmegen, Medical Centre

Thesis co-supervisor
Dr. P.J. Schaap
Assistant professor, Laboratory of Systems and Synthetic Biology
Wageningen University

Other members
Prof. dr. S.C. de Vries, Wageningen University
Prof. dr. T.R.J.M. Desmet, Ghent University, Belgium
Dr. S.A.F.T. van Hijum, Radboud University Nijmegen
Dr. R. de Jong, DSM Biotechnology Center, Delft

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

From existing data to novel hypotheses

Design and application of structure-based

Molecular Class Specific Information Systems

Remko Kuipers

Thesis

submitted in fulfilment of the requirements for the decree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. dr. M.J. Kropff,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Wednesday December 12th, 2012
at 11 a.m. in the Aula.

Remko K.P. Kuipers

From existing data to novel hypotheses. Design and application of structure-based Molecular Class Specific Information Systems.

232 pages

Thesis, Wageningen University, Wageningen, The Netherlands (2012)

With references, with summaries in Dutch and English

ISBN: 978-94-6173-350-4

Table of Contents

General Introduction

1.1 Life Sciences

The life sciences encompass a wide range of research fields aiming to understand biological systems on all levels. Today, four billion years after life first arose on earth, it has diversified into three domains: eukaryotes, prokaryotes, and archaea, each consisting of innumerable species. Several million species are currently found in nature and described by scientists and a multitude of this number has yet to be categorized [1]. An even larger number of species have evolved, thrived and become extinct again in the four billion year history of live on earth [2]. Life can be studied in many different ways, and while biology remains at the heart of the life sciences, many more specialized fields have sprung up. Life science research projects are involved in the study of organisms, understanding of biological processes and pathways, interactions within cells, transport of compounds within and between cells, the production of useful products, defence versus external threats, reactions to stimuli and stress, and many other areas of interest.

To be able to study life from all these different angles scientists produce and consume large amounts of data. One data driven classification for research fields in the life sciences is directly linked to the data aspect of biological systems that is studied. Nomenclature for these classifications commonly uses the subject of interest appended with the suffix omics, a neologism referring to obtaining a class of bio-data in high throughput. A large number of different omics techniques and methodologies have been developed of which a few that we encountered while working on this thesis are listed here.

1) Genomics;
2) Transcriptomics;
3) Proteomics;
4) Secretomics;
5) Interactomics;
6) Variomics;
7) Metabolomics;
8) Fluxomics;

Different aspects of biological systems are thus studied by different high-throughput omics techniques. Genomics projects, for example, focus on the genome of an organism. Transcriptomics focuses on the subset of those genes that are transcribed under specific conditions such as a disease or external stimuli. Proteomics studies the expression of proteins and secretomics how a subset of these proteins is secreted which is important for commercial protein engineering. Interactomics aims to understand the interactions among the many biological components in a system, for instance, the full protein-protein interaction network of a cell. Variomics studies the variation in DNA and protein sequences between individuals and populations for applications in healthcare. Metabolomics studies metabolites and fluxomics their fluxes in, out, and through cells or cell organelles.

The overall goal of research projects in all these fields is to improve the understanding of life in all its varieties. A broader viewpoint encompassing data from multiple fields and in-depth analyses are often required to study biological components and mechanisms. Small perturbations in biological processes may have huge consequences requiring studying the entire system instead of just a single component. Systems biology studies how higher-level properties emerge from complex networks of interactions among the many individual components of biological systems. Systems biology projects include an iterative cycle of experimental generation of heterogeneous data sets, the analyses of these data sets, and their subsequent integration into mathematical models. Several cycles are usually required to generated and fine tune a model for a specific purpose. The work described in this thesis mainly focusses on data integration. In addition, several tools are discussed for data analyses and in some projects predictions were experimentally verified.

Systems biology projects require high quality data to accurately predict biological phenomena. Nowadays, a substantial part of life science research is data centric requiring a wide variety of equipment, methodologies, and tools used to produce those bio-data. Biological data is thus produced by lots of different researchers, laboratories, and organizations, and made available to the research community in lots of different ways. Due to the size and complexity of these data sets, structured storage methods are required to allow research projects to access and utilize the full potential of these data sets. On-line databases are nowadays commonly used to offer public access to data from research projects. For example, institutes such as NCBI, EBI, or CMBI offer public access to databases that contain many different data types such as protein sequences [3–5], structures [6,7], annotations [3,8] or publications [9].

Frequently used data types, such as sequences and structures, have largely been internationally standardized. These data types tend to get deposited in publicly accessible databases hosted by large universities or governmental organizations that have the required financial backing to maintain these repositories for years on end. These data types, for example protein sequences or gene expression data, are generally of a low information content, and relatively easy to produce with currently available high throughput techniques. Higher quality, experimentally derived data sets such as pathway fluxes and ligand binding studies are more detailed. These datasets are, however, more difficult and expensive to produce because much more human expertise and more expensive equipment is required to generate the data. Public databases are often not available for these highly specialized data types due to the smaller number of producers and consumers. These data sets are therefore usually offered by the producers themselves, despite the costs and burden of maintaining in-house data repositories. New developments and improvements in equipment and software tools will lead to easier and more cost efficient methods of producing these datasets. Additionally, increased storage efficiency and an improved ability to store diverse data types in generic databases will lead to easier and cheaper hosting of data sets. Combined, these developments will lead to a continuous increase in the amount and quality of data that will become publicly available [10–12].

The availability of many diverse data sets for scientist is important as novel insights into systems of interest can be gained not only by novel research and experiments but also by combining and integrating existing data. Even data originally generated for a completely different purpose might be relevant and useful for a novel question when properly integrated with all other existing knowledge. Research projects thus produce and consume lots of different data types that have to be interlinked and integrated into a database to be of any use. This thesis describes how molecular class specific information systems (MCSISes) can be used to handle a vast amount of highly heterogeneous data and how it can use these different data types to answer a wide variety of protein structure and function related molecular questions. MCSISes implicitly and explicitly use the 4 billion years of evolution when coping with multiple structure and multiple sequence alignments as vehicles to transfer information between biological entities.

Heterogeneous data

Organizing multiple biological datasets into a single framework for use in research projects is often complicated by the large heterogeneity of those datasets. The combination of heterogeneous datasets, however, can be extremely beneficial for research projects. The following four examples demonstrate the usage of systems that integrate highly heterogeneous datasets and the application of these systems in biomolecular research projects. The first example comes from a project at the University of Nijmegen using an MCSIS, the subsequent three are from our own work with the 3DM platform, which is based on MCSIS technology.

1) Drug design projects focus on the discovery and development of medicines. Common targets for drug design studies are Nuclear Receptors, a class of transcription factors regulated by hormones or metabolites. Approximately 13% of all FDA drugs approved up to 2006 target a member of the Nuclear Receptor family, making it the second most often targeted family after the G protein-coupled receptors [13]. Nuclear receptors are composed of two main domains, a DNA binding domain (DBD) and a ligand binding domain (LBD), and a series of smaller domains of unknown structure and function. The DBD binds specific hormone response elements (HRE) on the genome which can be found in the promoter regions of genes. The promoter region of a gene can contain multiple HREs and an HRE can be present in the promoter of multiple genes. A single receptor can therefore regulate the transcription of a large number of genes and have a significant effect on the organism. Nuclear Receptors themselves are regulated by ligand binding to the LBD which activates or deactivates the receptor complex. The LBD region consists of an alpha helical sandwich containing binding sites for the ligand, co-activator, and co-repression molecules [14]. The human genome encodes 48 different nuclear receptors. Most of these receptors are associated with specific activator and inhibitor ligands. Other receptors are not strongly associated with a particular ligand but may instead be activated by metabolic intermediates such as fatty

acids, thereby acting as metabolic sensors. For several so-called orphan receptors however, specific ligands have not yet been found [15]. Nuclear Receptors can be activated or inhibited by conformational changes in the LBD which either blocks or opens up the ligand and co-activator binding sites. Hormone receptor activators and inhibitors are highly interesting for pharmaceutical companies either as medicine or to influence hormone related processes such as fertility, birth control, or certain types of cancers. Developing a new activator or inhibitor for a Nuclear Receptor is however a very time consuming and labour intensive process. The common procedure for drug target discovery is to create a large library of potential ligands. New leads coming from this library are then tested, modified, and improved using directed evolution until suitable candidates are found. These candidates are then tested first on animals, and when successful applied in small scale medical trials. The entire process of developing a new compound takes years and costs billions of dollars without the guarantee of a successful outcome. To better understand the underlying mechanisms of nuclear receptors and to aid in drug target development the NuclearDB molecular class specific information system (MCSIS) was developed [16]. The NuclearDB focuses on nuclear receptors from human and many other species and integrates data from different fields into a single database. Using large sequence alignments from the NuclearDB, Folkertsma *et al.* [17,18] discovered a residue position that is involved in inhibitor binding but not activator binding. This knowledge is crucial for a better understanding of the binding mechanisms, functional properties, and conformational changes of Nuclear Receptors, and can be directly applied to develop new activators and inhibitors.

2) Enzymes are capable of converting a wide range of substrates through a large number of different reactions. Some enzymes have specialized in the rapid conversion of a specific substrate whereas other enzymes are capable of converting a wide range of related substrates often at slower conversion rates. Substrate specificity and optimization of substrate conversion by enzymes is required for the improvement of existing or development of novel biotech processes such as for the production of chemicals. Bio-based methods, for example for the production of plastics or bio-fuels, are a fast growing market. Novel methods to replace existing chemical processes are also of interest as the chemical methods often produce pollution and require significant energy inputs.

In this example MCSIS technology is used to pin-point a key residue for substrate specificity in oxaloacetate hydrolase. Oxaloacetate hydrolase (OAH) is an enzyme that hydrolyses oxaloacetate, a product of the TCA cycle, into oxalic acid. Oxalate, the dianion of oxaloacetate, is a toxic compound often produced by fungi for self-defence purposes. The presence of the OAH gene can therefore be used as an indication of oxalate production and thus toxicity. Aspergillus niger, a fungi used for the production of food additives such as citric acid, contains the OAH gene and four additional homologous loci that may or may not be OAHs. It is important to determine whether A. niger, or any other commercially used

fungus, has the capability to produce the toxic oxalate compound and if so which gene(s) are responsible.

A 3D Molecular Class Specific Information System (3DM) for the Phosphoenolpyruvate mutase/Isocitrate lyase superfamily was built to store and connect available heterogeneous data sets. Study of correlated mutations between alignment positions in the superfamily revealed a network of nine residues [19]. For most of these positions the function was unclear, but several are located around the active site of the enzyme. Combined analysis of the correlated mutations network with protein annotations pointed out a serine that is completely conserved in a subset of proteins known to provide OAH functionality. Mutation studies were performed to study the role of this serine in potential OAH producing enzymes. Serine residues were mutated to other amino acids in proteins with known OAH activity and inserted into proteins without OAH activity. The results of these studies show that the serine is indeed required for the conversion of oxaloacetate through binding of the OH side group in the active side. The results also show that these mutations have hardly any effect on the affinity of OAH for any of its other substrates. Replacing the serine in the active site changes the conformation of the active site and reduces the binding potential of the conversion intermediate. The mutations therefore affected substrate affinity for one substrate only, and left the enzyme activity unchanged. Using correlated mutations and protein annotations stored in a superfamily knowledge base thus lead to the discovery of a key residue in the conversion of oxalocacetate.

A parallel study targeted L-galactono-γ-lactone dehydrogenase (GALDH) an enzyme part of the vanillyl-alcohol oxidase superfamily that consists of both dehydrogenases and oxidases. A strong correlation was found between an alignment position and the oxidase/dehydrogenase classification. Most oxidases have either a proline or a glycine residue on the position while dehydrogenases favored other residues. Mutating the alanine on the position to a glycine in GALDH led to a 400 fold increase in oxygen reactivity. Detailed study of the structure of the dehydrogenase indicated the targetted position acted as a gatekeeper, with larger residue preventing oxygen from reaching the active side. Mutating the residue did not change the active side and therefore did not affect the natural reduction potential of the GALDH enzyme [20].

3) Enzyme reaction mechanisms can be governed by several external factors. For example, the availability of co-factor and substrate, regulation of genes, and energy and carbon balances are all factors that influence the substrate conversion rate of an enzyme. To replace chemical processes by biological processes, for example for the production of high value chemicals, novel systems must either be cheaper, more environmentally friendly, or faster than existing methods. Improvement of enzyme activity is therefore another major focal point of enzyme engineering studies. The previous example showed the application of heterogeneous datasets to alter the substrate specificity of an enzyme, without changing

the activity. Using another superfamily the opposite was also shown to be possible: increase activity without changing the substrate affinity [19]. The RmlC like cupin superfamily consists of over 2,000 proteins which share a common ß-barrel fold. Cupin family members are functionally very diverse, though most members are enzymes with active site residues located each time at corresponding positions in the ß-barrel [21]. Correlated mutation analysis revealed a co-evolution network of positions mostly surrounding the active site. The pair of strongest correlating positions, interestingly, was not found near the active site but in a structurally conserved loop on the protein's surface. Little was known about the function of the residues in this network, and available literature for the superfamily listed relatively few single mutants at these positions. The available mutations in the two correlating positions in the surface loop almost universally lead to a decrease in activity, suggesting both positions are involved in substrate conversion. The combined amino acid occurrence data of both positions in a superfamily alignment of several thousand proteins revealed that five residue combinations were highly overrepresented (fig. 1.1). These five residue combinations, out of the possible 400, are present in more than 5% of the sequences, and these five combinations account for 65% of all sequences in the superfamily. The phosphoglucose isomerase (PGI) enzyme from Pyrococcus furiosus (PfPGI) was selected for mutational screening as it is one of the best characterized members of the superfamily. PfPGI has the combination PY as wild type in the conserved surface loop which occurs in only 0.78% of all superfamily members. Several single and double mutants were made at these positions and the enzyme activities of all mutants were determined. The single mutants both had reduced enzyme

%	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	-
A	0,12	0	0,18	0,12	0	**11,23**	0	0	0	0	0	0,06	0	0	0	0,18	0,06	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	0	0	0	0,12	0	0	0	0	0	0,12	0	0	0	0	0	0	0	0	0
E	0	0	0	0	0	7,75	0	0	0	0	0,12	0	0	0	0	0,06	0	0,06	0	0	0
F	0,42	0	0	0	0	0,06	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	1,62	0	**25,95**	0,18	0	1,08	0,06	0	0	0,24	0	0,30	0	0	0	0,60	1,14	0	0	0,06	0
H	0	0	0	0	0	0,06	0	0	0	0	0	0	0	0	0,06	0	0	0	0	0	0
I	0	0	0,18	0	0	1,20	0	0	0	0	0	0,84	0	0	0	0	0,06	0	0	0	0
K	0,60	0	0	0	0	4,44	0	0	0	0	0	0,12	0	0	0	0,30	0	0	0	0	0
L	0	0	0	0	0	0,30	0,06	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,06	0	0	0	0	0
N	0,18	0,12	0,06	0	0,30	0,06	0,06	0	0	0,90	0,06	0,06	0	0,06	**10,57**	0,18	0,06	0,12	0	0,18	0
P	0	0	0	0	0,12	3,36	0	0	0	0	0	0	0	0	0	0,06	0,06	0	0	0,78	0
Q	0,18	0	0	0	0	3,48	0,06	0	0	0,12	0	2,22	0,06	0	0	0,12	0	0	0	0,06	0
R	0,12	0	0,06	0	0	**10,27**	0,06	0	0	0	0	0,24	0	0	0,06	0	0	0	0	0	0
S	0	0	0	0	0	0,72	0	0	0	0	0	0,06	0	0	0	0	0	0	0	0	0
T	0,24	0	0	0	0	0,06	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V	0	0	0,12	0	0	1,74	0	0	0	0	0	1,38	0	0	0	0,12	0	0	0	0	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	0	0	0,06	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
-	0	0	0	0	0	0,24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,42

Figure 1.1: Partial residue occurrence data for two alignment positions. The 20 common amino acids are shown on both axes, and the occurrence of the corresponding amino acid pair is shown in the table itself. Abundant combinations are highlighted in bold. For example, 25,95% of all sequences in this alignment contain a GD motif on the two positions. The 0.42% of sequences containing no residues on both positions is a combination of alignment errors and truncated fragments.

activity (P27A resulting in AY that is not observed in superfamily members, and Y28G resulting in PG that is observed in 3.36% of superfamily members). Of the three double mutants, two were able to compensate for the loss of activity of the single-mutants. The third double mutant (P27A/Y28G leading to AG that is observed in 11.23% of superfamily members) not only compensated for the loss of activity caused by the single mutants, but increased activity to twice that of the wild-type enzyme. In summary, a large, high-quality protein alignment combined with correlated mutation data was instrumental in this project. Automatic scanning of available superfamily literature provided information regarding the possible roles of the positions of interest in the PfPGI enzyme. Protein activity could be improved without affecting substrate affinity because the surface-loop is far away from the active site residues in the folded structure. The next step is now, of course, to understand why this remote loop is so important for activity.

4) Detection of pathogenic non-synonymous mutations in disease related genes is a major focus of DNA diagnostics. In this example an MCSIS was used for a case study on pathogenicity prediction. Fabry's disease is a monogenetic disorder linked to the alpha-galactosidase (GLA) gene that was the selected target. Deficiencies in GLA lead to the accumulation of globotriaosylceramide in cells throughout the body. This accumulation leads to impairment of cell function in tissue throughout the body resulting in a highly diverse spectrum of symptoms. Correctly diagnosing patients for Fabry's disease without DNA diagnostics is therefore difficult [22]. Since Fabry's disease is monogenentic by nature, mutations in the GLA gene can potentially be used in clinical diagnosis. A mutation in the GLA gene is a difference between the GLA gene of a suspect individual with respect to the reference sequence. With on average a difference of 1 in every 1,200 nucleotides between two human individuals, distinguishing between natural variation and pathogenic mutations is a complex task as many differences between sequence data from patients and reference genomes are simple cases of natural variance.

Mutations are often classified into three groups based on phenotype: benign, malignant, and undetermined. Logically, malignant mutations are the focus of many research groups in hospitals around the world [23]. Statistically relevant and otherwise characterised malignant mutations found with DNA diagnostics based patient studies are often published in literature. Benign mutations can be retrieved from single nucleotide polymorphism (SNP) databases such as dbSNP [24]. The so called 'undetermined variants' however, pose a problem due to normal genetic variation between individuals. A method was developed to predict the pathogenicity of these variants using a wide variety of data types such as: the function of the mutated amino acid, endogenous function of the gene or protein, location in the protein structure, solvent accessibility, ligand contacts, existing mutational data, etc. [25]. To store the large amount of data required for these analyses a 3DM system was used for easy interlinking of data and rapid availability. Classifying variants for use in diagnosis

requires a high level of certainty because the results of a misdiagnosis range from a costly but ineffectual treatment to patient death. Variant classification should therefore rely on as many, high-quality data as possible.

Feature	*Value*	*Pathogenicity*	*Weight*
Occurrence of A at position 89	0.81%	95.94%	100.00
Conservation of most conserved residue	43.96%	43.96%	30.00
Sidechain accessibility	0.4%	93.57%	40.00
Bumps	0	0.00%	25.00
Hydrophobicity Switch	no	NA	NA
Buried hydrophilic position	no	NA	NA
Fabry disease mutation(s)	20	95.00%	100.00
Pathogenic mutations(s) from other human genes	7	70.00%	70.00
Identity of nearest sequence containing the mutant as wildtype	0.13	80.00%	50.00
Grantham distance	96	50.00%	10.00
Blosum62 score	-1	20.00%	10.00

Table 1.1: Statistical analysis of a mutation suspected to cause Fabry's disease mutation. Table shows the analysis for mutation p.L89A in GLA. Analysis based on the alpha-amylase superfamily that contains 4.704 sequences. Each part of the analysis is shown as a single line. For example, the first line shows that only 0.81% of the superfamily sequences contain an alanine at position 89. The pathogenicity and weight of the analysis are based on this value, and used for the overall pathogenicity prediction. Overall pathogenicity for this mutant was predicted as 77.6% likely pathogenic.

An alpha amylase protein superfamily, containing GLA, was used for variant classification. The superfamily contains almost 5.000 proteins and 3.700 known mutations from literature obtained using the Mutator tool (described in chapter 1.4). Additionally, thousands of annotations such as signal peptides, sequence conflicts, and active sites were retrieved from GenBank and Uni-Prot. Contact, RMSD, and solvent accessibility of alignment positions were determined using crystal structures. All available data are combined and weighed to calculate a pathogenicity factor to indicate the likelihood of the mutation being pathogenic (table 1.1). Validator includes a classifier to cluster known mutations and predict the pathogenicity of these mutations based on all these data. This knowledge can be used to determine whether observed mutants and phenotypes are indeed linked. Comparison with existing mutation databases, such as the HGMD, showed Mutator was able to collect more mutations from literature. Several mutations classically associated with Fabry's disease were evaluated using models of GLA in combination with the available data types. Evidence for pathogenicity could be found for most variants such as high conservation of the wildtype

residue in the superfamily or steric clashes due to sidechain orientation. Validator is however, not the final solution for patient diagnosis as the function and regulation of a gene can be influenced by many different factors and even subtle influences can have large repercussions. Validator was shown to give a clear overview of the effects of mutations in the gene; knowledge that can be used in clinical diagnosis.

Interoperability of Heterogeneous Data Types

The examples in the previous section illustrate the importance of collecting and linking heterogeneous data types, and show the difficulties encountered while doing so. To be of maximum use in research projects, data needs to be integrated into a framework of knowledge, be directly available, easy to use, and extendible with new knowledge. The research projects described in the examples use knowledge bases containing heterogeneous datasets interlinked with structure-based multiple sequence alignments. Heterogeneous data types can be linked when an inherent relation between the different types exists. For several life science data types such shared characteristics are present. For example, the link between sequence and structure data for a protein is straight forward. The nucleotides on the chromosome code for amino acids, which fold into the three-dimensional structure of the protein. Other data types such as pH resistance, viral immunity, or gene synteny are much harder to link as these data types do not describe similar or directly related biological entities. Life-science data is data of biological systems and should therefore preferably be integrated on the basis of a biological relationship. Sequence and structure alignments are therefore a suitable kernel to link many different data types, such as protein structures and residue annotations. Data can also be linked to other data using descriptions of the data and the relations between the data types. New metadata data models such as RDF [26] are currently implemented to describe a data type and the type of relation between one data type and another. These techniques are very flexible and applicable to almost all types of data. Biological knowledge can be incorporated into such a descriptive scheme using ontologies [27]. Many biological ontologies exist, each specializing in a different area such as transcriptomics or patient and disease related data. RDF based techniques are very flexible and can be used to link almost any data. Unfortunately datasets annotated using different ontologies are seldom interoperable resulting in many highly specialized applications that cannot be used for wider research.

A large number of bioinformatics tools are available to support researchers in life science projects. These tools attempt to offer an integrate environment for the storage, linking, and retrieval of heterogeneous data sets based on alignments, RDF annotations, or other techniques and frameworks. Several prominent types of bioinformatics tools are described in more detail in the next section. As the number of available tools is huge and specialized reviews are available aimed at parts of the bioinformatics eco-system [28–30] a complete overview is omitted.

Data Integration Tools

Access to many databases and tools is nowadays provided by webservices [31,32]. Webservices allow data to be retrieved from external sources, and tools to be run offsite using standardized protocols. This greatly simplifies the analysis process as the tools and databases used do not have to be installed locally. Furthermore, the standardized data format and communication protocols ensure tools written in different programming languages or on different operating systems can communicate with each other seamlessly. Many different webservices are available for example to retrieve sequences, predict phosphorylation sites, construct multiple sequence alignments, determine transcription factor binding sites etc. [33,34]. Several tools are available to combine these webservices into workflows to automate common procedures. Applications such as Taverna [35] and web-based platforms such as Mobyle [36] and REMORA [37] are examples of tools that offer the integration of webservices into workflows.

Reproducible Research Systems (RRS) is a relatively recent term used to describe systems that focus on the reproducibility of results by storing and publishing workflows of data analyses [38]. An RRS offers a persistent workspace were data can be stored and analysed using different tools. Workflows consisting of a combination of tools, data analyses, and parameter settings can be created and applied to different data sets. Workflows can be annotated and described extensively and offered publically to be used as supplementary materials in articles. RRS thereby tries to make it easier for third parties to perform the experiments themselves. RRS systems can use either webservices or tool-specific components for the actual analyses. Examples of RRS platforms are Galaxy [39] and GenePattern [40].

Model-driven development uses application models to generate software tools for custom purposes. These custom applications are built from stock components for data analysis, data import and export, and graphical reporting. A database is generated based on the datatypes described in the model, and a frontend is generated to enter, process, and visualize the data. Custom functionality can generally be included using external plugins or added using custom code in the model. Examples of these tools are MOLGENIS [41] and MEMOPS [42]. These tools are aimed at bioinformaticians rather than life science researchers as they require in-depth knowledge of techniques such as databases, XML, application models, and programming languages. The assembled applications however, can be hosted as web-application for general use.

The NuclearDB, described in the first example on heterogeneous data sets, is an example of a Molecular Class Specific Information System (MCSIS). MCSISes are knowledge bases that can be used to integrate and store heterogeneous data sets for a protein superfamily. MCSISes use structure-based multiple sequence alignments to interconnect different data sets. Reporting tools, data analyses, and data export options are all integrated in these systems. Links too many external databases are stored to facilitate data exchange. MCSIS systems are available for several protein families such as the Nuclear Receptors and G

Protein-Coupled Receptors. These systems were used to validate the scientific and technical concepts underlying MCSIS databases [16-18,43-47].

Many other tools exist to support life science research. Some offer combinations of the described techniques while others use completely different working models. geWorkbench [48], for example, integrates different functionalities into a workflow for later processing. However, only geWorkbench modules can be used for analyses of data, and those modules are executed on a grid. This allows for computationally intensive workloads to be designed by researchers for subsequent execution on one of the supported life science grids. Many other combinations of tools, analyses, and workflows are however, also possible and thousands of applications, tools, and services are available to support researchers in their work.

Each type of tool was designed for a particular aim, use-case, and target audience, and has associated advantages, and drawbacks. Webservice based tools are aimed at researchers with some background in bioinformatics. Creating workflows in these programs is not unlike programming and some expertise in data handling is therefore required. The black-box model of webservices where tools and databases are not available locally however results in several drawbacks. The inner workings of webservices, in- and output formats, and functional parameters are often sparsely documented if any documentation is available at all. Additionally, though the communication protocols used between service and client are standardized, the data formats are not. Converters are therefore often required before data from one service can be passed to another. The reliability, availability, and performance of webservices also vary as the services are offered by third-parties. Webservices for instance may stop working or disappear completely without prior notice.

RRS systems and MCSIS databases are aimed at researchers and therefore designed to be intuitive and user-friendly. Installation and maintenance by experienced users however remains required to keep the data up-to-date. RRS systems additionally often only offer tools installed locally for analysis and data manipulation. These systems are designed to be reproducible and are therefore often hesitant to depend on external factors. Users of RRS systems are therefore unable to take advantage of the wide range of functionality available through webservices as the continued availability of webservice cannot be guaranteed. RRS systems therefore run the risk of being used only halfway into a research projects after researchers have manually obtained and prepared there data. This workflow severely reduces the reproducibility of experiments. Model-driven development tools are aimed squarely at bioinformaticians. A lot of in-depth knowledge of software installations and data types is required to create the initial model and generate the application. Even with the large number of available tools, researchers with an informatics background often create programs of their own or mix and match different applications to facilitate research projects. These tools use many different techniques to achieve a wide variety of goals. However, these applications and workflows are often designed to solve one particular problem, and are therefore highly specialized for the research field and project they originate from.

This thesis focusses on the development and extension of the 3DM platform. The aim was the creation of an easy to use web-based application that offered the integration of many different datasets and could be applied to any protein superfamily. Tools that combine generic data handling and specific biological knowledge, such as MCSISes, are applicable to many different projects and therefore highly useful to assist researchers in many different fields. 3DM was based on the MCSIS technology developed at the Centre for Molecular and Biomolecular Informatics, Radboud University Nijmegen, Medical Centre. MCSISes can be used for vastly diverse projects by using a combination of structure and sequence based alignments to link available data sets. Import and export options of data allow MCSISes to be used in research projects in conjunction with webservices and other technologies. My work has focused on the development of the knowledge base system itself, the underlying database, and several tools incorporated into the 3DM platform. I have been personally involved as main programmer and system designer in the 3DM projects used as examples above.

Evolutionary Considerations

3DM and MCSIS databases are centred on protein super families, and use structure-based multiple sequence alignments to interconnect datasets. Sequence alignments are used as base or kernel to connect the various data types as they reflect the outcome of millions of years of evolution. Evolution, having been defined as 'descent with modification', nevertheless manifests itself in conservation on several levels. Protein regulation and function in biological systems are most conserved as significant changes in either property often lead to reduced fitness of the organism. Protein structure and sequence are much less conserved as the structure or sequence is irrelevant as long as the regulation and function of the protein are conserved. However, the preservation of regulation and function often leads to conservation of structure and to a lesser extent sequence. Proteins can be clustered into superfamilies using these conservation characteristics. These superfamilies consist of large groups of evolutionary related homologs from many different organisms all with similar roles and functions [49].

Proteins are the active components in many biological systems and they provide valuable functionalities such as gene translation, substrate conversion, viral immunity systems, energy production, and many more. The function of a protein is determined by a combination of structure, sequence, and dynamics. Protein structure and sequence are subject to evolution, and so are thus the dynamics and ultimately the function. Chromosomal mutations and rearrangements continuously adapt the genomic sequence, which results in changes to the sequence, structure, and function of proteins. The function of proteins however, is preserved through evolutionary pressure, as defunct proteins often lead to competitive survival problems for the whole organism. Similar functionality can, however, be provided by different genes as a result of gene duplications, horizontal gene transfer, or unbalanced chromosomal translocations. These duplicated genes are much less restricted by evolutionary pressure as only one of the homologs is required for the primary

function. After gene duplication events some homologs are therefore free to diverge in sequence, structure, and function. Due to their common evolutionary origin these proteins initially remain structurally related until consecutive mutations leads to divergence beyond recognition. Genes that have been duplicated before branch points in the phylogenetic tree of life have ended up in multiple organisms through speciation. Superfamilies, therefore can contain multiple genes from multiple organisms coding for related but separable functions. Superfamilies thus reflect the structural, functional and evolutionary relationships between proteins.

Homologous Residues

During evolution proteins are conserved primarily on function, and less on structure and sequence. Conserved regions in the structure or sequence of a protein are therefore often also functionally conserved [47]. Residues at conserved positions have evolved from an ancestral residue at the same position, similarly to how genes in related organisms have evolved from a single gene in the common ancestor. These homologous residues are therefore often involved in similar functions (fig. 1.2). Consequently, information can be carried over between homologous residues; any functional annotation found for a residue in one protein is likely also applicable for the corresponding residue in a homologous protein. Determining which residues are homologs within a superfamily is therefore an important step in the conversion from data to knowledge. Types of annotations that have been shown to be transferable between homologous residues include ligand contacts, active site annotations, salt bridges, a role in mobility, phosphorylation, etc. Data can however only reliably be transferred between truly homologous residue positions. Alignment quality thus is more important than the size of the alignment.

1.2 Alignments

Sequence Alignments

Sequence alignments, normally, are performed by programs such as ClustalW [50] and MUSCLE [51] that compare sequences one amino acid pair at a time. A Dayhoff style substitution matrix (fig. 1.3) is used to assign a score to each amino acid pair. To construct a matrix an overview of allowed mutations within a range of sequences must be obtained and evaluated. Originally, matrices were derived from large manually obtained sequence alignments for groups of highly similar sequences. The prototype amino acid substitution matrix is the PAM matrix calculated by Dayhoff in 1978 using alignments of proteins that share 85% identity. PAM matrices are however relatively unsuitable for aligning evolutionary more distantly related proteins due to the short evolutionary distance covered by the alignment used to calculate the matrix. Therefore, more sophisticated substitution design strategies were devised and used to create newer matrices. Henikoff & Henikoff [52], for example, used alignments of protein blocks to generate the BLOSUM matrices.

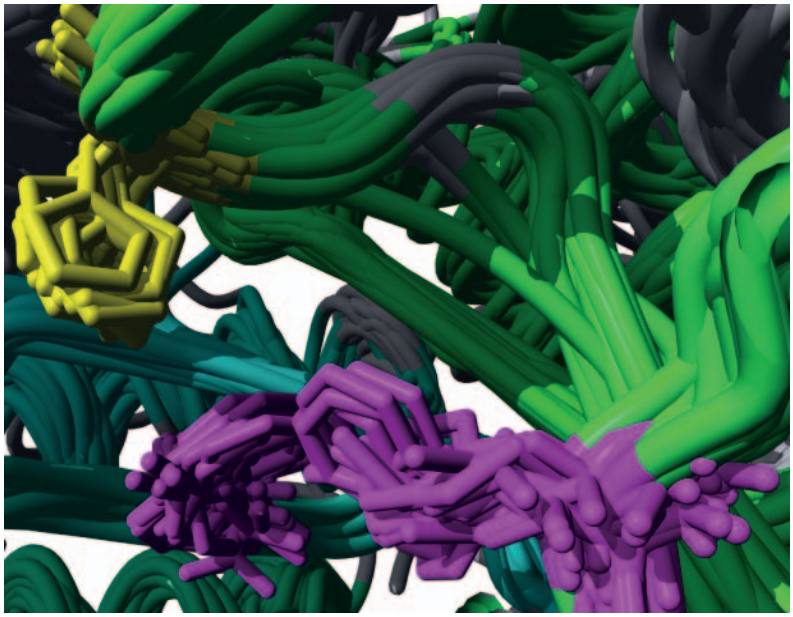


Figure 1.2: Superposition of 29 structures from the alpha-amylase superfamily with three homologous positions highlighted. The structures are shown in tube style with structurally conserved regions shown in green and variable regions in grey. The scene contains one conserved position (yellow) and two correlated positions (magenta). The conserved position is a histidine in 96.1% of the 4,900 aligned sequences. The two correlated positions are the strongest correlating pair found in the correlation network of this superfamily. Three amino acid combinations, WF, EW and QT, account for 60% of the sequences.

C Cys	12																		
S Ser	0	2																	
T Thr	-2	1	3																
P Pro	-3	1	0	6															
A Ala	-2	1	1	1	2														
G Gly	-3	1	0	-1	1	5													
N Asn	-4	1	0	-1	0	0	2												
D Asp	-5	0	0	-1	0	1	2	4											
E Glu	-5	0	0	-1	0	0	1	3	4										
Q Gln	-5	-1	-1	0	0	-1	1	2	2	4									
H His	-3	-1	-1	0	-1	-2	2	1	1	3	6								
R Arg	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6							
K Lys	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5						
M Met	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6					
I Ile	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5				
L Leu	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6			
V Val	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4		
F Phe	-4	-3	-3	-5	-5	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9	
Y Tyr	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10
W Trp	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W

Figure 1.3: The PAM 250 matrix by Dayhoff.

In principle, substitution matrices contain for each of the 20 proteogenic natural amino acids the odds of either a mutation to any of the other 19 amino acids, or that the residue remains conserved. Substitutions that result in small changes in amino acid size, charge, and/or hydrophobicity are more likely to be accepted than more invasive changes and this is reflected in the substitution matrices. In the Dayhoff matrix shown in figure 1.3 for example, tryptophan has non-negative substitution scores only for tyrosine and phenylalanine which are large aromatic amino acids like tryptophan. Similarly, the highest scoring substitution for lysine is arginine which preserves the positive charge of the side chain. The odds of residue conservation are also incorporated into the matrices by observing the number of synonymous versus non-synonymous mutations. Amino acids that are often involved in structural or functional properties, such as cysteine and tryptophan, therefore score higher for synonymous mutations than residues such as serine and alanine.

Substitution matrices are used to generate an optimal pairwise alignment between two sequences. Alignments can be either global or local depending on the alignment algorithm. Local alignment methods such as BLAST search for regions of similarity between two sequences, global alignments attempt to align the two sequences completely. Local alignments are therefore better suited for more distantly related sequences that share regions of sequence similarity while global alignments are more suited for closely related sequences. To create a local alignment of two sequence regions the two sequences are placed on the x and y axis and for each position the substitution score for the two amino acids is entered into the matrix. The most optimal alignment is determined by walking through the matrix along the path that gives the best score (fig. 1.4). Optimal global alignments often contain gaps to properly align two sequences. The example alignment matrix in figure 1.4 has two solutions depending on whether gaps are allowed or not. If gaps are allowed, and the gap open penalty does not exceed -1, the highlighted path is the highest scoring solution. If gaps are not allowed, or the gap open penalty exceeds -1, a direct path aligning K-C and E-N

	P	G	F	K	N	L	P	L	E	D	Q
P	6	-1	-5	-1	-1	-3	6	-3	-1	-1	0
A	1	1	-5	-1	0	-2	1	-2	0	0	0
F	-5	-5	9	-5	-4	-5	-5	2	-5	-6	-5
C	-3	-3	-4	-5	-4	-6	-3	-6	-5	-5	-5
E	-1	0	-5	0	1	-3	-1	-3	4	3	2
L	-3	-4	-5	-3	-3	6	-3	6	-3	-4	-2
P	6	-1	-5	-1	-1	-3	6	-3	-1	-1	0
L	-3	-4	-5	-3	-3	6	-3	6	-3	-4	-2
D	-1	1	-6	0	2	-4	-1	-4	3	4	2
D	-1	1	-6	0	2	-4	-1	-4	3	4	2
Q	0	-1	-5	1	1	-2	0	-2	2	2	4

Figure 1.4: Dayhoff derived alignment matrix for the alignment of two amino acid sequences. Residues of sequence one are shown in the left column and residues of sequence 2 in top row. The best alignment is determined by the highest scoring path starting from top left to bottom right.

omits gaps and scores higher. Gap open, and gap elongation penalties can thus be used to force the algorithm to favour single larger gaps, multiple smaller gaps, or no gaps at all.

Several matrices exist to determine the alignment between related sequences such as BLOSUM [52], PAM [53], and BATMAS [54]. Most of these matrices are available in multiple versions depending on the evolutionary divergence of the sequences used to construct the matrix. For example, the BLOSUM62 matrix was derived using an alignment of sequences at least 62% identical to each other, and the BLOSUM80 matrix derived from sequences at least 80% identical. The BLOSUM62 matrix is therefore better suited for the alignment of sequences that have diverged more in comparison to the BLOSUM80 matrix as BLOSUM62 includes substitution from nature on larger evolutionary distances. However, for alignments of closer evolutionary divergence a BLOSUM80 matrix is better suited as the BLOSUM62 matrix will allow substitutions that are unlikely to occur at close evolutionary distances. The final goal of sequence alignment procedures is a so-called Multiple Sequence Alignment (MSA). An example is shown in figure 1.5 where 31 nuclear hormone receptors are aligned by both ClustalW and 3DM.

Using matrices has several disadvantages. Main drawback is that most matrices are themselves based on alignments of sequences or sequence blocks, and thus the choices made when constructing those alignments have a direct impact on the quality of the matrix. Any bias or error in the original alignment will directly translate into a biased matrix which in turn leads to biased alignments. Many authors have suggested improved methods for the design of substitution matrices. Jones *et al.* [55] start with an alignment of sequences that are clustered at the 85% sequence identity level. From this alignment a Dayhoff-style matrix is generated by multiplying the matrix by itself that then can be used for the alignment of family members with a lower sequence identity. Ng and Henikoff made more generic family specific substitution matrices like, for example, a matrix for membrane proteins [56]. Overington *et al.* [57] designed matrices for buried and accessible residues, for helical and strand residues etc. These matrices can be used when at least one structure is available from which the characteristics of each position can be determined. This method was generalised for fold recognition (threading) by Bowie et al [58].

Hogeweg and Hesper [59] further generalised the alignment technology by introducing profile based sequence alignments. Profiles are based on multiple (aligned) input sequences and each column in a profile reflects the allowed amino acid variability at the corresponding position in the alignment. In a sense, a profile can be seen as one matrix for each position in the alignment. In case of threading these matrices are determined from the characteristics of the protein structure, but in iterative profile based alignment procedures such as described by Jones et al, and as implemented in WHAT IF [60], the profiles are build-up from the growing alignment itself (see chapter 2.4).

In 3DM, we use a hybrid method based on structure superpositions and iterative profile alignments that are explained in the remainder of this chapter. The next section

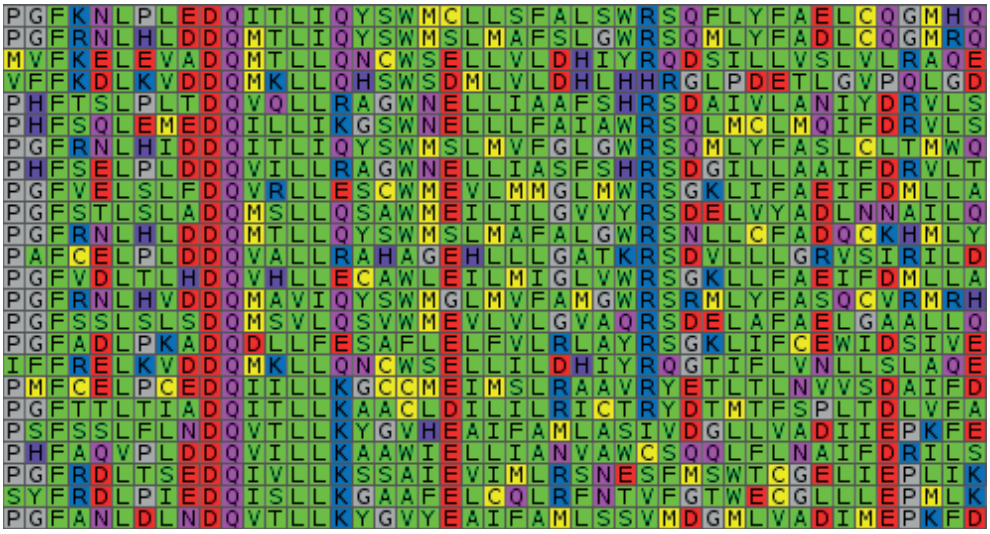
discusses structure superposition/comparison methods, and this is followed by the heart of 3DM, the structure-based sequence alignments.

Structure Superposition

Structures remain conserved over a longer evolutionary time than protein sequences. Consequently, for two or more distant homologous proteins it should be easier to correctly superpose the structures than to correctly align the sequences. This was already recognized in the famous article by Chothia and Lesk [61] on the relation between sequence and structure similarity. Sander and Schneider [62] quantified this relation in their seminal article in which also the HSSP database technology is introduced and iterative profile alignments are applied. Vriend and Sander [63]showed in 1991 that it is possible to automatically superpose protein structures, thereby opening up a whole field of research that in this thesis culminates in the 3DM platform. Structure superposition however is a CPU intensive problem that can be solved in a myriad of different ways. Hundreds of methods and algorithms are therefore available today to create structure alignments.

Superposing two structures can be defined as finding the best fit between the backbones of both structures. Changing the superposition to improve the match in one region of the superposition however often leads to a reduced fit in another region resulting in an overall non-optimal solution. To superpose two highly similar structures a least-squares fitting algorithm is traditionally used. These algorithms determine the optimal rotations and translations by minimizing the sum of the squared distances among all structures in the superposition [64]. More recently a wide range of different methods has been applied to superpose both highly similar and highly diverse structures. Rather than describing all programs and methods individually, this section will start with a description of the WHAT IF algorithm used by 3DM. Several other programs are subsequently discussed to illustrate some of the different methods available to superpose structures. A complete overview of the field is difficult to obtain, however a series of reviews and papers together with the Wikipedia entry on structural alignment can be used as a starting point [64-67].

WHAT IF [63] uses a three-step process to create a pair-wise superposition. During the first step superposable fragments from two structures are determined by comparing all fragments of 10 to 15 residues in length using a distance metric. A list of superposable fragments is generated containing all pairs of fragments for which the C-alpha atoms are within a specified distance cutoff. Each fragment pair is then superposed using a least-squares algorithm. The superposition of each fragment is evaluated using the root mean square distance and maximum distance between C-alpha atoms. If these distances do not exceed cutoffs of 2.0Å and 3.8Å respectively the fragment pair is considered to be superposable. Superposed fragments are subsequently elongated by adding one residue to the C-terminal end of both fragments after which the two fragments are again superposed. The elongation process continues until the addition of new residues to the fragments results in RMSD



```
PGFKNLPLEDQITLIQYSWMCLLSFALSWRS----Q--FLYFA-----------ELCQGMHQ
PGFRNLHLDDQMTLIQYSWMSLMAFSLGWRS----Q--MLYFA-----------DLCQGMRQ
MVFKELEVADQMTLLQNCWSELLVLDHIYRQ----DSILLV-------------SLVLRAQE
VFFKDLKVDDQMKLLQHSWSDMLVLDHLHHR-GLPDET---------LGVPQLGD-------
PHFTSLPLTDQVQLLRAGWNELLIAAFSHRS--DA----IVLA----------NIYDRVLS-
PHFSQLEMEDQILLIKGSWNELLLFAIAWRS------QLMCLM----------QIFDRVLS-
PGFRNLHIDDQITLIQYSWMSLMVFGLGWRS----Q--MLYFA-----------SLCLTMWQ
PHFSELPLDDQVILLRAGWNELLIASFSHRS--DG----ILLA----------AIFDRVLT-
PGFVELSLFDQVRLLESCWMEVLMMGLMWRS--G----KLIFA-----------EIFDMLLA
PGFSTLSLADQMSLLQSAWMEILILGVVYRS--D----ELVYA-----------DLNNAILQ
PGFRNLHLDDQMTLLQYSWMSLMAFALGWRS----N--LLCFA-----------DQCKHMLY
PAFCELPLDDQVALLRAHAGEHLLLGATKRS--DV----LLLG----------RVSIRILD-
PGFVDLTLHDQVHLLECAWLEILMIGLVWRS--G----KLLFA-----------EIFDMLLA
PGFRNLHVDDQMAVIQYSWMGLMVFAMGWRS----R--MLYFA-----------SQCVRMRH
PGFSSLSLSDQMSVLQSVWMEVLVLGVAQRS--D----ELAFA-----------ELGAALLQ
PGFADLPKADQDLLFESAFLELFVLRLAYRS--GK---LIFC------------EWIDSIVE
IFFRELKVDDQMKLLQNCWSELLILDHIYRQ----GTIFLV-------------NLLSLAQE
PMFCELPCEDQIILLKGCCMEIMSLRAAVRY--ET----LTLN-----------VVSDAIFD
PGFTTLTIADQITLLKAACLDILILRICTRY--DT----MTFS-----------PLTDLVFA
PSFSSLFLNDQVTLLKYGVHEAIFAMLASIVDGLL---VA--------------DIIEPKFE
PHFAQVPLDDQVILLKAAWIELLIANVAWCS----QQ--LFLN-----------AIFDRILS
PGFRDLTSEDQIVLLKSSAIEVIMLRSNESF--MS---WTCG------------ELIEPLIK
SYFRDLPIEDQISLLKGAAFELCQLRFNTVF--GT---WECG------------LLLEPMLK
PGFANLDLNDQVTLLKYGVYEAIFAMLSSVMDGML---VA--------------DIMEPKFD
```

Figure 1.5: Partial alignment of 31 Nuclear Receptors by 3DM (top) and ClustalW (bottom). The 3DM alignment shows the alignment positions at the top, amino acid colouring based on chemical properties. Each line contains the sequence of one protein. Several positions show different characteristics: complete conservation on positions 29 and 30, charged residues on position 17, hydrophilic positions such as 15 and 16, and positions with a seemingly random distribution such as 3 and 26. Structurally variable regions as identified by 3DM have been removed from both alignments.

or maximum C-alpha distances exceeding the thresholds. Fragments that contain residues already included in other superposed fragment pairs are ignored to prevent overlapping fragments. Additionally, fragments with less than four non-helical residues are ignored as helices tend to always be superposable. The second step of the superposition process consists of a clustering of superposed fragments to determine if the fragments are part of a larger common structure. The distances between the centers of mass and rotation matrix of the fragments are compared to exclude fragments that cannot be clustered together. The final step of the procedure is a fine-tuning and pruning of the clusters. This step is required as several shortcuts and optimizations are used to speed up the superposition resulting in a lower overall accuracy. The largest cluster of superposed fragments is again superposed to minimize the distance between equivalent C-alpha atoms. The list of equivalent residues is then re-examined to ensure all equivalent C-alpha atoms are within the maximum distance of each other and are part of a stretch of equivalent residues. To create a superfamily superpositioning 3DM uses WHAT IF to create pair-wise alignments of all superfamily structures with a master template. From the superfamily superposition WHAT IF produces a sequence alignment of structurally equivalent residues that can be used to fine-tune a common structural core.

To create a multiple structure alignment many programs use a progressive approach starting from a series of pair-wise alignments. The pair-wise alignments can be created by a variety of methods and are stored for easy comparison. MALECON [68] for example creates a library of pair-wise structure alignments using SoFist [69]. A multiple-structure alignment is generated by progressively combining the pair-wise alignments into one overall alignment. Another example of the progressive alignment approach is Mustang [67] that uses dynamic programming for the pair-wise alignments.

MUSTA [70], MultiProt [71], and MASS [72] use geometric hashing procedures to find superposable fragments. Geometric hashing procedures start by creating a list of spatial coordinates for both objects. These objects are then superposed by randomly selecting pairs of coordinates and transforming one of the objects onto the other object. If a transformation results in additional matching coordinate pairs the superposition is accepted. MASS searches for fragments present in all structures whereas MultiProt and MUSTA support fragments only superposable in a subset of the structures. Several techniques are used to elongate the fragments, and to combine fragments with a highly similar transformation to create an overall superposition. These methods create a multiple-structure alignment using both combinatorial and pair-wise procedures and are similar to the procedure used by WHAT IF.

Ye and Godzik [73] developed a method called Partial Order Structure Alignments (POSA), to superpose structures using partial order graphs (POGs) and preserve structurally variable regions in the visualization. A guide tree is used to align multiple structures pair-wise using an iterative procedure. POSA alignments focus more on including all structure variation into the superposition as opposed to WHAT IFs focus on determining structurally equivalent residues.

CE-MC [74] uses pair-wise structure comparison using a combinatorial expansion (CE) algorithm as input. A Monte Carlo (MC) optimization technique is used to create random changes in the alignment between two structures. A distance-based scoring matrix is used to reject or accept the changes. Several types of permutations are applied both backwards and forwards to the aligned fragments in the MC procedure. Validation of the results on manually curated families shows an increase in alignment positions and an increase in alignment distance over the initial CE derived alignments. Due to the Monte Carlo analysis and pair-wise comparisons CE-MC is relatively slow.

Structure superpositions are highly useful and suitable for a wide range of applications. Creating a superposition can however be time consuming especially when many or highly divergent structures are involved. Therefore, several databases are available that contain predetermined superpositions or classifications of structures into families based on these superpositions. SCOP [75], CATH [76], and DALI's FSSP database [77] are examples of these databases. These three methods work different in most details, but have in

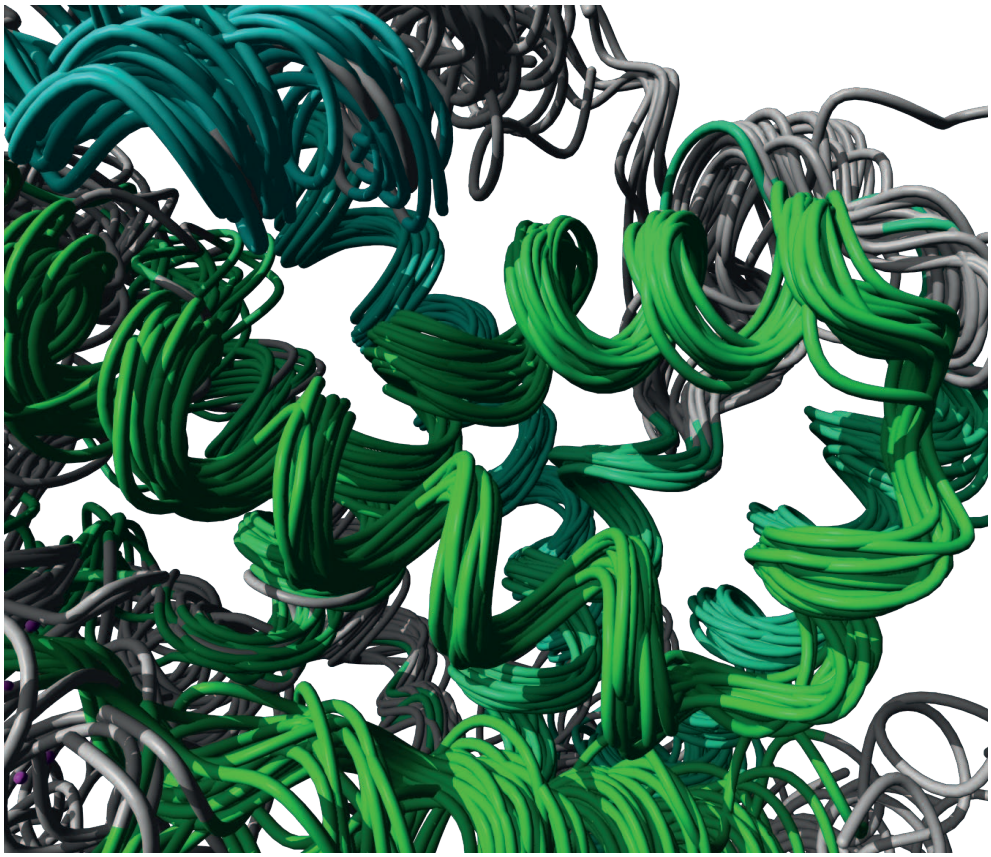


Figure 1.6: Superposed helices in 26 structures from the NR superfamily. Regions in green/cyan are part of the common structural core, grey regions are structurally variable.

common that all structures available from the PDB are structurally classified into a family hierarchy. Structure classifications from these databases combined with structure superpositions and sequence alignments can be used to assemble large evolutionary divergent structure-based multiple sequence alignments for protein superfamilies.

The first step in the 3DM procedure to create a superfamily database is retrieving all structures classified as belonging to the superfamily. From these superfamily structures a superposition is generated (fig. 1.6). The superposition can then be used to determine the common structural core of the superfamily. Sequence alignments using the structurally conserved regions as anchors can be used to include sequences for which no structure is available. As long as structures of homologous proteins are used for the determination of the common structural core, very high quality alignments can be obtained. These alignments are therefore perfectly suited to construct a superfamily alignment and to interconnect heterogeneous data sets. The combination of structure superpositions and iterative profile based sequence alignments used to create 3DM databases is detailed in the next section.

Structure-Based Sequence Alignments

For 3DM superfamily systems we combine structure and sequence alignments into a single superfamily system (fig. 1.7). To cover the entire family an initial superfamily wide superposition is created by superposing all superfamily structures on a master template (fig. 1.7a1). A common core of structurally equivalent residues is determined based on all superpositions (fig. 1.7a2). Template structures are selected for subsequent sequence alignments based on sequence similarity. A profile based iterative sequence alignment consisting of four rounds is used to align sequences to the templates (fig. 1.7b[1-4]). The common core residues are used as anchors in the sequence alignments. A superfamily alignment is assembled from the individual subfamily alignments. For each aligned sequence the residues in the common core are determined and used in the superfamily alignment (fig. 1.7c). A technical description of the alignment process and implementation details can be found in chapter 2.4.

Using a combination of structure superpositions and sequence alignments has the advantage that proteins can be included in the alignment based on either structure or sequence similarity. This results in a much lower detection limit for distantly related proteins. Structure is more conserved then sequence, and the templates used for the sequence alignments are at most 80% identical. 3DM is therefore able to include sequences into a superfamily database that in a pair-wise alignment would have a sequence identity below 5% while still being reasonably confident that the common core residues are properly aligned. Compared to structure or sequence alignments, a structure-based sequence alignment thus broadens the data included into the alignment and can improve the quality of the sequence alignment.

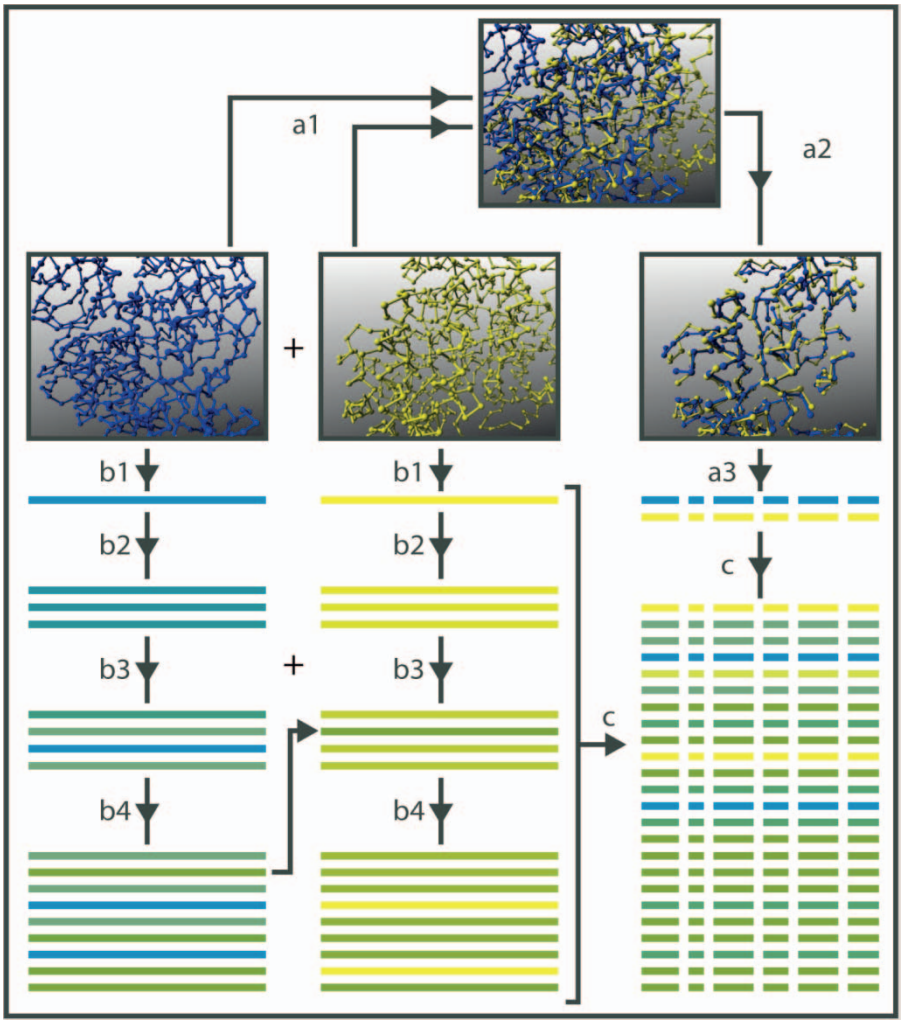


Figure 1.7: 3DM MSA generation procedure. A superfamily superposition is generated using all structures from the superfamily (a1). From the superposition the common structural core of the superfamily is determined (a2). Structural templates are selected, and the common core residues of these templates are determined (a3). Individual sequence alignments are generated for each template using an iterative profile base alignment procedure (b). The common core residues of each aligned sequence are determined by comparison to the common core of the templates. The resulting alignments are combined into a superfamily alignment.

Within 3DM, alignments are used as basis to incorporate diverse data sets and are used as a primary key to link these data sets. For example, mutations, contacts, and annotations can be interconnected using alignments. The remaining sections of the first chapter will outline some of the different datasets available within 3DM systems.

1.3 Correlated Mutations

Datasets consist of many variables some of which might hold dependencies. A dependency between two or more variables indicates a possible correlation between the two variables. Two variables are dependent upon each other when a change in variable A can be used to predict the change in variable B and vice versa. In enzymes, for example, residues in the active site are often closely correlated with the substrate. However, many other factors also contribute to substrate binding specificty, and there is not necessarily a direct correlation between active site residues and the substrate. However, any correlation between residues and substrates can be used for various purposes such as predictions of possible substrates for newly sequenced proteins or for analyses of the function of active site residues. Many examples of such correlations have been found in life science data. The following sections will describe several examples of correlations and highlight some of the major contributions to the field.

Correlation Based Structure Prediction

Correlations can be determined within and between diverse biological datasets. Correlations can for example be determined for expression data with disease state using microarrays, and protein reaction efficiency with bound catalysts using essays. One of the first computational examples of correlations in the life sciences was RNA structure prediction. RNA structures are formed by base-pairing of nucleotides in the RNA molecule (fig. 1.8). Parsch *et. al* [78] determined correlations between complementary regions of RNA molecules using alignments and used these regions to predict RNA structure. Stretches of matching nucleotides were used to predict loops and multiple matching stretches were combined to predict the 2D conformation of the entire RNA molecule. Perfect conservation of nucleotides in the alignment was not required, however preservation of base-pair combinations was. Therefore, an A-T pair could be replaced with a C-G, T-A, or G-C pair but C-T, T-C, G-A, and A-G combinations were not allowed on the positions as this would prevent base-pairing. An example application of this method is shown in figure 1.8. The nucleotides on positions 10 through 16 match the nucleotides on positions 19 through 25, and two more regions of nucleotide correlations can be observed. These data can be used to predict the loops formed between these nucleotides. The conformation of the complete RNA molecule shown in figure 1.8 can be modelled using the combined predictions of the loops.

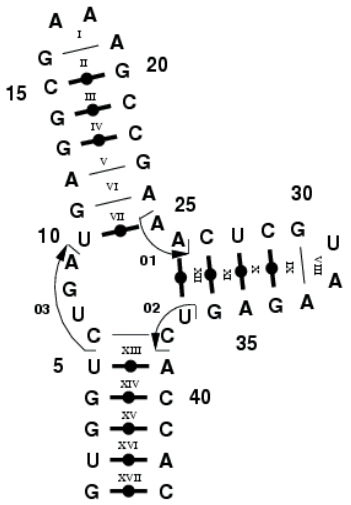


Figure 1.8: 2D structure of an RNA molecule.

The link between residues or sequence fragments is less direct for proteins. Besides cysteine bridges, no unambiguous direct amino acid contacts are formed in proteins. For instance, a tryptophan does not inexorably 'bind' to a tyrosine as a cytosine does to a guanine. Protein structures are therefore harder to predict than RNA structure. Several attempts have been made, for example by Göbel *et al.* who tried to determine secondary structure using correlations to predict intra-protein contacts [79]. Their method was tested on 11 protein families and validated using known structures belonging to the families. The results showed a shorter average distance between the C-beta atoms of amino acids in predicted pairs versus random amino acid pairs. The prediction accuracy of their method however ranges between 40% and 70% which is insufficient for high-quality structure predictions. More recently, Marks *et al.* used a correlation method and data from multiple sequence alignments to predict which residue pairs would be in close proximity in folded structures [80]. 3D models were generated with constraints placed on the distance allowed between residues predicted to be in proximity of each other. In total, 15 proteins with lengths between 50 and 260 amino acids were modelled using this technique. For each model the RMSD error of at least two-thirds of the residues fell within a 2.7-4.5Å band of the actual structure. This technique is therefore very promising but improvements in RMSD and an increase in the length of the sequences studied are required before applications for full protein modelling are possible.

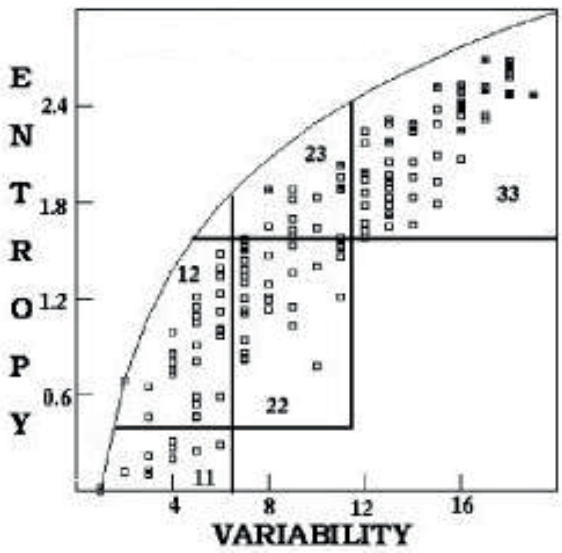


Figure 1.9: Entropy variability plot. Residues are divided into five boxes: 11) main functional site, 12) core residues surrounding the main functional site, 22) core residues potentially involved in structural or signaling roles, 23) residues interacting with modulators, and 33) surface residues without a defined function.

Correlations Between Life-Science Datasets

Oliveira *et al.* used the amino acid variability and entropy obtained from alignments to predict the function of amino acids on alignment positions [47]. Five alignments consisting of several hundreds of proteins were created for the Globin chains, Ras-like, and Serine-

proteases families. These families were selected as lots of studies have been done on individual family members, and the function of most residues is therefore known. The variability and entropy of each position in the alignments was plotted in an entropy-variability graph (fig. 1.9). Using family specific criteria the residue positions were assigned to boxes and functional predictions were made for each box. The residues found in the main functional sites of the proteins were predicted to fall in box 11. The predictions matched overall with the known function for the residues. Additionally, positions bordering two boxes often showed functional properties of both boxes.

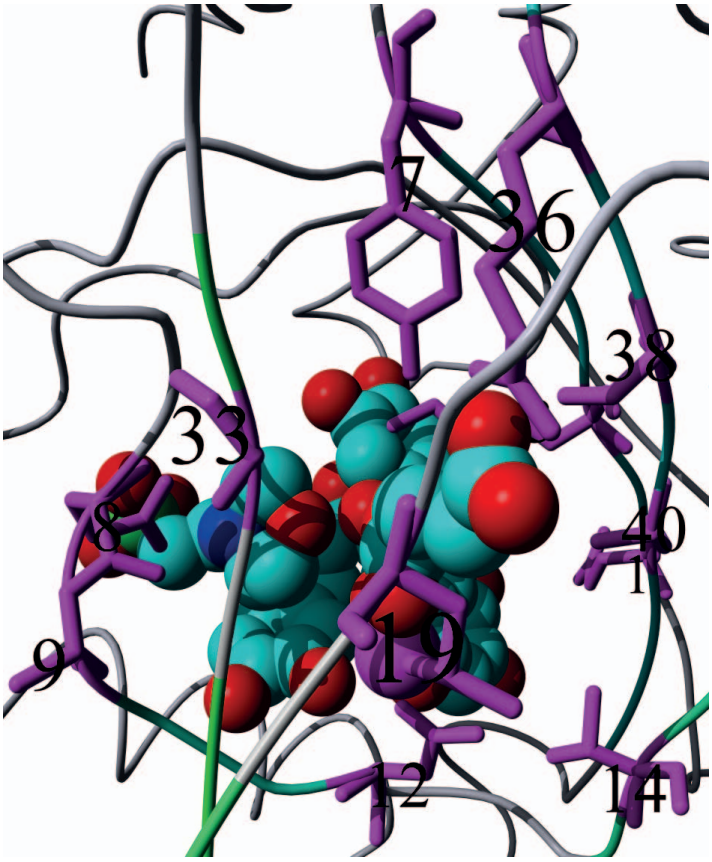


Figure 1.10: Network of correlated positions surrounding the active site of an enzyme from the cupins superfamily. Scene shows structure 1GP6 with 4 ligands and an iron ion bound to the active site. Positions in the correlation network are shown in magenta, with sidechain atoms, and labeled using alignment numbering. Positions in the common core are green/cyan, and positions in variable regions are grey. Some variable residues have been removed for clarity.

Correlations Between Alignment Positions

Alignments reflect the evolutionary history of a superfamily. Some residue positions within an alignment are unconstrained and can probably be mutated without any detrimental effect on the protein's function (fig. 1.5, positions 3 and 26). Other residue positions however, are either strictly conserved or correlated and linked with specific protein functions. Correlated positions must be mutated in tandem to prevent loss of function. This applies to for example positions involved in charge clamps, substrate binding, or dimer interfaces. Evolutionary pressure is thus visible in alignments as correlations between such alignment positions (fig. 1.10). These correlations between alignment positions can be calculated using the distributions of amino acids on both positions. Single mutants on positions involved in substrate binding or recognition are often disruptive. Double or triple mutants however, may change the layout of the pocket sufficiently for a different ligand to be recognized. Many residues on these positions interact and must be changed simultaneously (either experimentally or by natural processes) for a protein to recognize a different ligand. Positions involved in a similar function thus often cluster in correlated networks (fig. 1.11). Correlation networks between alignment positions can therefore be used to study groups of functionally related positions in alignments, and predict the function of positions in a cluster.

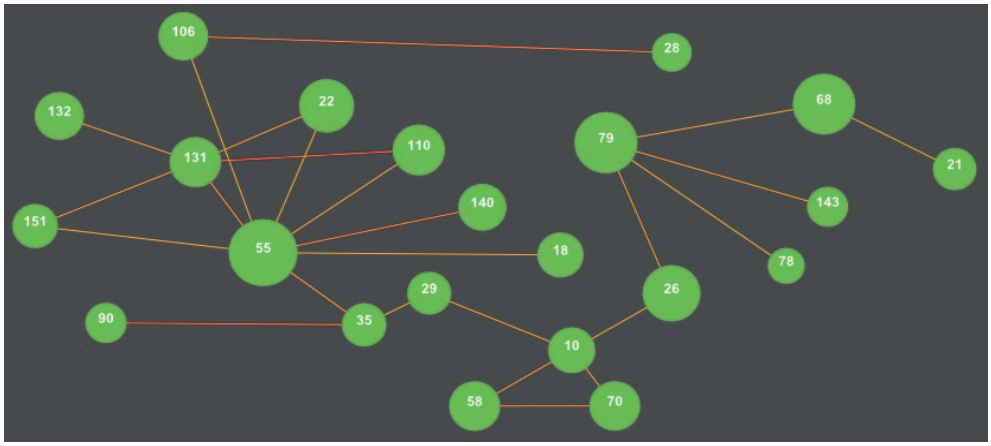


Figure 1.11: Correlations in the Nuclear Receptor superfamily. Alignment positions shown as nodes, pair-wise correlations above 0.7 are shown as edges. Note that several alignment positions correlate with more than one alignment position. Examples are 10, 55, 79, and 131.

Algorithms

Several algorithms are available to determine correlations between datasets for 3DM we use the statistical coupling method [81]. Other popular algorithms include mutual information method [82,83], perturbation method [84], and Pearsons correlation method [79].

Lockless and Ranganathan [81] hypothesized that functionally coupled positions in an alignment can be determined by studying amino acid distributions. The distribution of amino acids on each position of the alignment is expected to approach the mean amino acid abundance determined for all proteins. When two positions in an alignment are functionally coupled, the evolutionary pressure to preserve this coupling will be visible in the correlation of amino acid distributions on both positions. The statistical coupling energy of two positions was determined by comparing the statistical energy vector at both positions in the entire alignment with the energy vector of the two positions representing a perturbation of amino acid frequencies on one of the two positions. To test their hypotheses, Lockless and Ranganathan used a structural alignment of the PDZ domain superfamily. Four structures were available of distantly related family members that were structurally well conserved with an average RMSD for Cα atoms of 1.4Å but sequentially only 24% identical. An active site position linked to ligand specificity was selected for the analysis, and the statistical coupling was determined between the selected position and all other positions in the alignment. Only 10 positions in the alignment showed coupling with their position of interest, 6 of these positions were either structurally near the position of interest, or known to be part of the interaction surface. The other 4 positions were assumed to be involved in signalling throughout the molecule. A series of double mutants on their position of interest with selected coupled and non-coupled positions in the protein yielded confirmed their results.

Types of Correlation and Potential Problems

Correlated mutations between alignment positions reflect an evolutionary pressure present in nature. Therefore the number of sequences in an alignment and the evolutionary distance between these sequences are indicative of the type of correlations that can potentially be extracted from it. For example, superfamily alignments containing evolutionary distant sequences are usually composed of proteins with a wide range of functions. The overall correlation signal that can be retrieved from these alignments therefore often points at residues essential for protein activity. More closely related alignments focussed on a subfamily of proteins with a known related functions show correlations angled towards (substrate) specificity or binding affinity.

Correlations are mathematical expressions of dependence between datasets, or variables within datasets. Most correlation algorithms assume the datasets and variables are normally distributed. Calculating correlations on biased datasets will result in correlation artefacts which appear as genuine correlations. For correlations between alignment positions this translates into a requirement of properly sampled alignments. For example, false positives correlations can occur when calculating the correlations for family consisting of two taxonomic branches using an alignment in which sequences belonging to one branch are heavily overrepresented. Conversely, properly constructed alignments have shown that including additional sequences into the alignment has no major effects on the correlations. Regenerated 3DM systems have shown for multiple families that the addition of new

structures and sequences does not fundamentally change the topology of the network of correlated mutations. This highlights several important aspects of both correlations and alignments: 1) alignments reflect actual evolutionary relationships between proteins; 2) Correlated mutation analyses on superfamily alignments are not influenced by small alignment errors; 3) The method used by 3DM to assemble superfamily alignments is robust, repeatable, and produces high quality alignments.

1.4 Mutation Information

Amino acid residues in a protein structure are involved in many different functions such as folding, signalling, ligand binding, etc. Mutating these residues can potentially result in protein deficiencies such as improper folding, missing hydrogen bonds required for ligand binding, or introduction of charged residues in the hydrophobic core of the protein. These deficiencies often lead to disruptive phenotypes and the effects of the mutations can therefore be used to study the role of specific amino acids positions in proteins. Effects of mutations can also be transferred between proteins using alignments. The effect of mutations in a protein of interest can therefore be predicted using mutations in homologs. It is therefore of vital importance to construct large libraries of mutations when studying protein and amino acid function.

Availability of Mutations

Mutations are available from two main sources: mutation databases and scientific publications. Manually curated mutation databases exist for various (human) proteins of interest, such as for the TP53 tumour repressor protein [85,86] and the BRCA1 and BRCA2 breast cancer genes [87]. Several databases focussed entirely on a single species are also available such as the HGMD [88] and PMD [89] for human disease mutations and the MGD [90] for mutations in mice. These databases are mostly updated and curated manually, and it is therefore time consuming to keep them up-to-date. For example, when in 1995 PMD was first published it contained only data from articles published before 1992. PMD has been updated several times since 1995, lastly in 2007. The 2007 update however contained no data from articles published after 2003. Keeping hand-curated mutation database up-to-date will become increasingly more difficult as the publication rate of articles continues to increase. Automation of (parts of) the retrieval of mutations from literature is therefore required to continue to maintain these databases. Several problems arise during this automation that will be discussed in the remainder of this chapter.

Retrieving articles using a web browser is relatively straightforward. Retrieving articles for a superfamily however is much more complicated as this requires retrieving hundreds or even thousands of articles from many different websites. Potentially relevant articles can be preselected using PubMed searches. A search for articles using the keyword ‘cancer’ as query however, yields 2.6 million results. Adding the keyword ‘mutation’ to the query reduces the number of hits to 138 thousand articles. Carefully constructed search

queries are therefore required to balance between downloading thousands of articles and missing mutations because the article was ruled out by the query. Retrieving the articles themselves is also non trivial as articles are not available from a single source. The location an article can be retrieved from can be obtained from the PubMed website. The articles themselves are mostly hosted by publishers such as the Nature Publishing Group or ScienceDirect. There is however no standardized method or location on publisher websites from which the article can be retrieved. Therefore, publisher specific procedures are required which greatly complicates the automation of article retrieval.

The second major difficulty lies in the extraction of mutations from articles as there is no universally accepted nomenclature for writing mutations in articles. Some publishers and papers enforce nomenclature guidelines, however, these have not been standardized. Authors therefore, often use a field specific nomenclature which can also differ from person to person. An additional complication is the lack of protein or gene names explicitly associated to specific mutations in articles. A process called grounding is often required to determine the protein and amino acid using data from the article such as species and gene names. For a complete list of steps in the mutation extraction process and the difficulties encountered in each step see chapter 2.6.

Mutation Retrieval Tools

A multitude of tools are available to assist in the automation of one or more aspects of the mutation retrieval process. Most of these tools specialize in only one or two aspects of the retrieval process such as tagging of mutations in articles or retrieval of mutations for a specific gene family. For the 3DM platform we required a tool to automatically search and download articles related to a protein superfamily, find mutations in the article, link the mutations to the corresponding protein, remove false-positives, and store the results in a database. We designed Mutator [25] to collect mutations for a complete protein superfamily based on the concepts of MuteXt [43]. MuteXt was used as base because it was the most complete tool available when development of Mutator was initiated. Furthermore, MuteXt contains many of the required features such as automatic selection and retrieval of articles, usage of protein annotations, assignment of mutations to sequences, and superfamily support. Mutator has been applied to several superfamilies and has shown to be an excellent tool for the retrieval of superfamily mutations [19,20,25]. An overview of the individual steps of the mutation retrieval process, available tools, and potential pitfalls is available in chapter 2.6.

1.5 Genetic Disorders

Pathogenicity of mutations

One of the many applications of mutation data is in diagnosis of genetic disorders. Many disorders have been linked to one or more genes using techniques such as genome wide association studies. For example, Fabry's disease and the GLA gene [91], breast cancer and BRCA1 and BRCA2 [92], various cancers and TP53 [93]. One of the motivations behind the Human Genome Project was the potential applications of a complete human genome in healthcare. Sequencing individual human genes had been done for years, however the

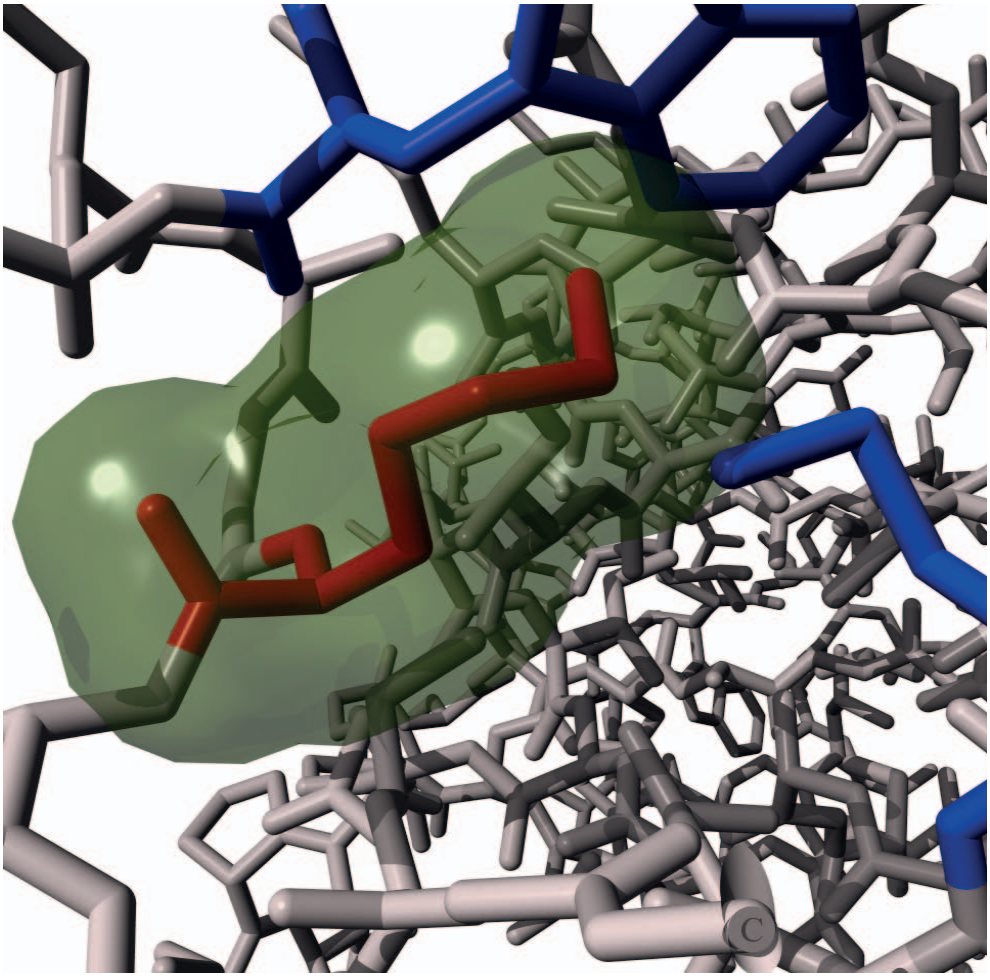


Figure 1.12: Validator scene of variant p.Y86K of hGLA. The variant is shown in red, with the Van der Waals radius in green. Steric hindrance is observed with residues W81 (top) and M296 (lower right), both highlighted in blue.

Human Genome Project was the first attempt to sequence and assemble the entire complex human genome. The project started in 1990 and was originally devised as a 15 year project. Advances in equipment and computing, and competition with commercial ventures however lead to a mostly complete release of the genome 2 years ahead of schedule [94,95]. The US part of the project was funded with 3 billion dollars, the total cost of the combined efforts of all project partners is estimated to be much higher. Nowadays both the time and finances required to sequence a full human genome have been reduced significantly through improvements in technology and methodology [96,97].

As sequencing costs go down, mass sequencing of disorder related genes for diagnostic purposes is becoming feasible [98]. Sequence data from both complete genomes, and specific disease related genes from patients, has revealed that natural variation in human genes is relatively common. Several projects are therefore currently underway to study and document the human variome: the genetic diversity found between individuals and between populations [99]. For human, on average 1 nucleotide in 1,200 differs between individuals, with an even larger genetic divergence between individuals from different ethnic backgrounds [100]. Differences between sequence data from a patient and the "reference" genome can therefore not be linked directly to pathogenicity as most variants are natural variations and have no significant direct effect on the health of carriers. For gene based disease diagnostics a method is thus required to distinguish between natural and pathogenic variants, and to predict the pathogenicity of potentially pathogenic variants. Validator (table 1.1 & fig. 1.12) was developed for the 3DM platform to predict the possible pathogenicity of unknown variants (UVs) found in patient data. Several publications are available describing the use of 3DM superfamily systems to predict the effects of mutations [19,20,25,101-103].

Validator

Validator uses the many different data types available from 3DM systems to predict the pathogenicity of UVs. Effects of the variant on protein structure can be studied using a visual representation of the variant, either from a PDB or a homology model (fig. 1.12). An accurate determination of the pathogenicity of a variant is very difficult due to the large number of factors involved. Amino acids are involved in protein folding, bonds involved in substrate recognition, catalytic residues, and many other roles. Mutations of amino acids can therefore result in a large number of different phenotypes associated with genetic disorders.

Due to the difficulties involved in pathogenicity determination Validator is not the final step in diagnosing genetic diseases, but merely an aid in the diagnosis of patients. Usage of Validator in combination with other techniques can lead to substantial improvements, and personalization of treatments which benefits both doctors and patients. Data types used by Validator and alternative tools are discussed in chapter 2.7.

1.6 Applications

This thesis deals with 3DM systems and their potential and application in research in the life sciences. Structure-based multiple sequence alignments augmented with external data and statistical analyses, can be of great assistance to life science researchers. The 3DM platform was designed to automatically construct 3DM systems for specific protein superfamilies. 3DM systems have been successfully applied to a wide range of applications in protein engineering, DNA diagnostics, and drug design both in industry and academia. They have been used to:

1) Study the functional mechanisms of a superfamily [103]

2) Improve substrate specificity [19] or thermo-stability [104]

3) Increase enzyme activity [19,20] or enantio-selectivity [103]

4) Predict the effects of mutations in disease linked genes [25]

The first two chapters of this thesis aim at providing a scientific and technical background for the 3DM platform. A general overview of the 3DM technology is provided in chapter 3. Project Hope, a mutant pathogenicity predictor, is described in chapter 4. These first 4 chapters focus mainly on the technical and scientific background and implementation of both 3DM and project hope. The focus in the next 4 chapters shifts from bioinformatics to applications of 3DM systems in research projects. Chapters 5 and 6 describe subsystems added to the 3DM platform: Comulator for correlated mutation analyses, Mutator for the retrieval of mutations from literature, and Validator for variant pathogenicity predictions. The final two chapters describe research projects in which 3DM was applied to study the alpha-beta hydrolase superfamily and enzymes in the alpha-amylase superfamily. The 3DM platform, project Hope, and the 3DM YASARA plugin are the three projects that I have devoted most of my time to. My contribution to the research projects described in chapters 7 and 8 has been limited to the design and implementation of algorithms and data facilities that enabled the bioinformatics aspects of these studies.

To complement the scientific background found in chapter 1, chapter 2 provides further insights in the inner workings of 3DM systems. This chapter is almost by definition incomplete because 100.000 lines of code (excluding the millions of lines of code in external software such as WHAT IF, YASARA, Utopia, etc.) cannot be described in detail in one chapter. It is only hoped that chapter 2 gives an impression of the scope of the software together with details on the implementation of concepts discussed in this introductory chapter.

Chapter 3 provides a more concise overview of 3DM and its evolution from earlier MCSIS systems. Thirteen 3DM systems were created and published online for examination. Examples from enzyme engineering, DNA diagnostics, and drug design projects highlight possible applications and uses of these public systems.

Chapter 4 describes Project Hope a tool to analyse protein mutations, similar to 3DM's Validator. Project Hope is aimed at biomedical researchers and relies on webservices, reusable external components, for data retrieval. A decision tree is used to predict the effects of the mutation on both the function and structure of the protein. Project Hope is developed and hosted at the Center for Molecular and Biomolecular Informatics (CMBI) at the Radboud Medical Center in Nijmegen as part of the PhD of H. Venselaar.

Chapter 5 describes the incorporation of the Comulator module for correlated mutation analyses (CMA) into the 3DM platform. A CMA algorithm based on statistical coupling was implemented to calculate CMA scores between alignment positions. Heatmaps, residue combinations on correlated positions, and correlation scores were integrated into the 3DM visualization modules. Four superfamily alignments of enzyme families were analysed using the CMA algorithm. For each of the four families a network of functionally related positions was found. Mutational studies showed that residues from these networks were mostly involved in specificity and activity.

Chapter 6 describes Mutator; the 3DM module for retrieval of mutation data from literature. A mutation database was assembled for the alpha-amylase superfamily and applied to Fabry disease. Validator was written as a DNA diagnostics tool and uses all data available in the 3DM platform to predict the pathogenicity of undetermined variants in disease linked genes.

Chapter 7 describes a 3DM system for the alpha-beta hydrolase superfamily. This family contains in excess of 15.000 proteins grouped into 5 distinct structural folds. Due to the size of the alignments several adaptations to the 3DM platform were required to be able to create superfamily alignments, facilitate the storage, and visualize the contents of the database. Automatic generation of phylogenetic trees was also added to the 3DM platform for this work. The alpha-beta hydrolase family contains numerous enzymes that have been, or could be targets for protein engineering. Subsequent publications by Jochens et. al [105], Jochens & Bornscheuer [106], and Hasenpusch et. al [107] describe protein engineering projects using the alpha-beta hydrolase 3DM system.

Chapter 8, finally, describes the application of a 3DM system in an enzyme engineering project focused on thermostability. Study of the enzyme's crystal structure, and rationally designed libraries from 3DM were used to increase the half-life of the enzyme at temperatures used in industrial processes.

References

1. C. Mora, D.P. Tittensor, S. Adl, A.G.B. Simpson, B. Worm, How Many Species Are There on Earth and in the Ocean?, PLoS Biology, 9 (2011) e1001127.
2. D. Raup, Biological extinction in earth history, Science, 231 (1986) 1528–1533.
3. M. Magrane, U. Consortium, UniProt Knowledgebase: a hub of integrated protein data, Database, 2011 (2011) bar009–bar009.
4. K.D. Pruitt, T. Tatusova, W. Klimke, D.R. Maglott, NCBI Reference Sequences: current status, policy and new initiatives, Nucleic Acids Res., 37 (2009) D32–36.
5. D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, E.W. Sayers, GenBank, Nucleic Acids Res., 39 (2011) D32–37.
6. P.E. Bourne, B. Beran, C. Bi, W.F. Bluhm, D. Dimitropoulos, Z. Feng, D.S. Goodsell, A. Prlić, G. B. Quinn, P. W. Rose, J. Westbrook, B. Yukich, J. Young, C. Zardecki, H.M. Berman, The evolution of the RCSB Protein Data Bank website, Wiley Interdisciplinary Reviews: Computational Molecular Science, 1 (2011) 782–789.
7. S. Velankar, Y. Alhroub, A. Alili, C. Best, H.C. Boutselakis, S. Caboche, M.J. Conroy, J.M. Dana, G. van Ginkel, A. Golovin, S.P. Gore, A. Gutmanas, P. Haslam, M. Hirshberg, M. John, I. Lagerstedt, S. Mir, L.E. Newman, T.J. Oldfield, C.J. Penkett, et al., PDBe: Protein Data Bank in Europe, Nucleic Acids Res., 39 (2011) D402–410.
8. E.C. Dimmer, R.P. Huntley, Y. Alam-Faruque, T. Sawford, C. O'Donovan, M.J. Martin, B. Bely, P. Browne, W. Mun Chan, R. Eberhardt, M. Gardner, K. Laiho, D. Legge, M. Magrane, K. Pichler, D. Poggioli, H. Sehra, A. Auchincloss, K. Axelsen, M.-C. Blatter, et al., The UniProt-GO Annotation database in 2011, Nucleic Acids Res., 40 (2012) D565–570.
9. E.W. Sayers, T. Barrett, D.A. Benson, E. Bolton, S.H. Bryant, K. Canese, V. Chetvernin, D.M. Church, M. Dicuccio, S. Federhen, M. Feolo, L.Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D.J. Lipman, Z. Lu, T.L. Madden, T. Madej, D.R. Maglott, et al., Database resources of the National Center for Biotechnology Information, Nucleic Acids Res., 38 (2010) D5–16.
10. EBI, UniProtKB/TrEMBL Release Statistics, UniProtKB/TrEMBL Release Statistics, (2011).
11. WWPDB, PDB: Yearly growth of total structures, Content Growth Report, (n.d.).
12. NCBI, NIH Manuscript Submission System Statistics, (n.d.).
13. J.P. Overington, B. Al-Lazikani, A.L. Hopkins, How many drug targets are there?, Nat Rev Drug Discov, 5 (2006) 993–996.
14. D.J. Mangelsdorf, C. Thummel, M. Beato, P. Herrlich, G. Schütz, K. Umesono, B. Blumberg, P. Kastner, M. Mark, P. Chambon, R.M. Evans, The nuclear receptor superfamily: the second decade, Cell, 83 (1995) 835–839.
15. M. Robinson-Rechavi, H. Escriva Garcia, V. Laudet, The nuclear receptor superfamily, J. Cell. Sci., 116 (2003) 585–586.
16. F. Horn, G. Vriend, F.E. Cohen, Collecting and harvesting biological data: the GPCRDB and NucleaRDB information systems, Nucleic Acids Res, 29 (2001) 346–349.
17. S. Folkertsma, P. van Noort, J. Van Durme, H.-J. Joosten, E. Bettler, W. Fleuren, L. Oliveira, F. Horn, J. de Vlieg, G. Vriend, A family-based approach reveals the function of residues in the nuclear receptor ligand-binding domain, J. Mol. Biol., 341 (2004) 321–335.
18. S. Folkertsma, P.I. van Noort, R.F.J. Brandt, E. Bettler, G. Vriend, J. de Vlieg, The nuclear receptor ligand-binding domain: a family-based structure analysis, Curr. Med. Chem., 12 (2005) 1001–1016.
19. R.K.P. Kuipers, H.-J. Joosten, E. Verwiel, S. Paans, J. Akerboom, J. van der Oost, N.G.H. Leferink, W.J.H. van Berkel, G. Vriend, P.J. Schaap, Correlated mutation analyses on super-family alignments reveal functionally important residues, Proteins, 76 (2009) 608–616.

20. N.G.H. Leferink, M.W. Fraaije, H.-J. Joosten, P.J. Schaap, A. Mattevi, W.J.H. van Berkel, Identification of a gatekeeper residue that prevents dehydrogenases from acting as oxidases, J. Biol. Chem., 284 (2009) 4392–4397.
21. J.M. Dunwell, A. Purvis, S. Khuri, Cupins: the most functionally diverse protein superfamily?, Phytochemistry, 65 (2004) 7–17.
22. C.L. Marchesoni, N. Roa, A.M. Pardal, P. Neumann, G. Cáceres, P. Martínez, I. Kisinovsky, S. Bianchi, A.L. Tarabuso, R.C. Reisin, Misdiagnosis in Fabry disease, J. Pediatr., 156 (2010) 828–831.
23. M. Erdos, K. Németh, B. Tóth, T. Constantin, E. Rákóczi, A. Ponyi, A. Dajnoki, J. Grubits, I. Pintér, F. Garzuly, K. Hahn, K. Bencsik, L. Vécsei, G. Fekete, L. Maródi, Novel sequence variants of the alpha-galactosidase A gene in patients with Fabry disease, Mol. Genet. Metab, 95 (2008) 224–228.
24. S.T. Sherry, M.H. Ward, M. Kholodov, J. Baker, L. Phan, E.M. Smigielski, K. Sirotkin, dbSNP: the NCBI database of genetic variation, Nucleic Acids Res, 29 (2001) 308–311.
25. R.K. Kuipers, H.-J. Joosten, R.H. Lekanne dit Deprez, M.M. Mannens, P.J. Schaap, Novel tools for extraction and validation of disease-related mutations applied to Fabry disease, Hum. Mutat, 31 (2010) 1026–1032.
26. X. Wang, R. Gorlitsky, J.S. Almeida, From XML to RDF: how semantic web technologies will change the design of "omic" standards, Nat. Biotechnol., 23 (2005) 1099–1103.
27. I. Spasic, S. Ananiadou, J. McNaught, A. Kumar, Text mining and ontologies in biomedicine: making sense of raw text, Brief. Bioinformatics, 6 (2005) 239–251.
28. R.D. Combes, In silico methods for toxicity prediction, Adv. Exp. Med. Biol., 745 (2012) 96–116.
29. C. Durrant, M.A. Swertz, R. Alberts, D. Arends, S. Möller, R. Mott, P. Prins, K.J. van der Velde, R.C. Jansen, K. Schughart, Bioinformatics tools and database resources for systems genetics analysis in mice--a short review and an evaluation of future needs, Brief. Bioinformatics, 13 (2012) 135–142.
30. M.H. Medema, R. van Raaphorst, E. Takano, R. Breitling, Computational tools for the synthetic design of biochemical pathways, Nat. Rev. Microbiol., 10 (2012) 191–202.
31. S. Pettifer, J. Ison, M. Kalas, D. Thorne, P. McDermott, I. Jonassen, A. Liaquat, J.M. Fernández, J.M. Rodriguez, D.G. Pisano, C. Blanchet, M. Uludag, P. Rice, E. Bartaseviciute, K. Rapacki, M. Hekkelman, O. Sand, H. Stockinger, A.B. Clegg, E. Bongcam-Rudloff, et al., The EMBRACE web service collection, Nucleic Acids Res., 38 (2010) W683–688.
32. M. Garcia, J. Karlsson, O. Trelles, Web service catalogue for Biomedical Grid infrastructure, Stud Health Technol Inform, 159 (2010) 76–87.
33. G. Benson, Editorial Nucleic Acids Research annual Web Server Issue in 2011, Nucleic Acids Research, 39 (2011) W1–W2.
34. G. Benson, Editorial Nucleic Acids Research annual Web Server Issue in 2010, Nucleic Acids Res., 38 (2010) W1–2.
35. D. Hull, K. Wolstencroft, R. Stevens, C. Goble, M.R. Pocock, P. Li, T. Oinn, Taverna: a tool for building and running workflows of services, Nucleic Acids Res., 34 (2006) W729–732.
36. B. Neron, H. Menager, C. Maufrais, N. Joly, J. Maupetit, S. Letort, S. Carrere, P. Tuffery, C. Letondal, Mobyle: a new full web bioinformatics framework, Bioinformatics, 25 (2009) 3005–3011.
37. S. Carrere, J. Gouzy, REMORA: a pilot in the ocean of BioMoby web-services, Bioinformatics, 22 (2006) 900–901.
38. J.P. Mesirov, Accessible Reproducible Research, Science, 327 (2010) 415–416.
39. J. Goecks, A. Nekrutenko, J. Taylor, T. Galaxy Team, Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences, Genome Biology, 11 (2010) R86.

40. M. Reich, T. Liefeld, J. Gould, J. Lerner, P. Tamayo, J.P. Mesirov, GenePattern 20, Nature Genetics, 38 (2006) 500–501.
41. M.A. Swertz, M. Dijkstra, T. Adamusiak, J.K. van der Velde, A. Kanterakis, E.T. Roos, J. Lops, G.A. Thorisson, D. Arends, G. Byelas, J. Muilu, A.J. Brookes, E.O. de Brock, R.C. Jansen, H. Parkinson, The MOLGENIS toolkit: rapid prototyping of biosoftware at the push of a button, BMC Bioinformatics, 11 Suppl 12 (2010) S12.
42. R.H. Fogh, W. Boucher, J.M.C. Ionides, W.F. Vranken, T.J. Stevens, E.D. Laue, MEMOPS: data modelling and automatic code generation, J Integr Bioinform, 7 (2010).
43. F. Horn, A.L. Lau, F.E. Cohen, Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors, Bioinformatics, 20 (2004) 557–568.
44. F. Horn, GPCRDB: an information system for G protein-coupled receptors, Nucleic Acids Research, 26 (1998) 275–279.
45. J.J.J. Van Durme, E. Bettler, S. Folkertsma, F. Horn, G. Vriend, NRMD: Nuclear Receptor Mutation Database, Nucleic Acids Res, 31 (2003) 331–333.
46. B. Vroling, M. Sanders, C. Baakman, A. Borrmann, S. Verhoeven, J. Klomp, L. Oliveira, J. de Vlieg, G. Vriend, GPCRDB: information system for G protein-coupled receptors, Nucleic Acids Res, 39 (2011) D309–319.
47. L. Oliveira, P.B. Paiva, A.C.M. Paiva, G. Vriend, Identification of functionally conserved residues with the use of entropy-variability plots, Proteins, 52 (2003) 544–552.
48. A. Floratos, K. Smith, Z. Ji, J. Watkinson, A. Califano, geWorkbench: an open source platform for integrative genomics, Bioinformatics, 26 (2010) 1779–1780.
49. A.E. Todd, C.A. Orengo, J.M. Thornton, Evolution of function in protein superfamilies, from a structural perspective, J. Mol. Biol., 307 (2001) 1113–1143.
50. M.A. Larkin, G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, J.D. Thompson, T.J. Gibson, D.G. Higgins, Clustal W and Clustal X version 20, Bioinformatics, 23 (2007) 2947–2948.
51. R.C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Research, 32 (2004) 1792–1797.
52. S. Henikoff, J.G. Henikoff, Amino acid substitution matrices from protein blocks, Proc. Natl. Acad. Sci. U.S.A, 89 (1992) 10915–10919.
53. M.O. Dayhoff, National Biomedical Research Foundation (U.S.), Atlas of protein sequence and structure Supplement, National Biomedical Research Foundation, Washington, D.C., 1973.
54. R.A. Sutormin, A.B. Rakhmaninova, M.S. Gelfand, BATMAS30: amino acid substitution matrix for alignment of bacterial transporters, Proteins, 51 (2003) 85–95.
55. D.T. Jones, W.R. Taylor, J.M. Thornton, The rapid generation of mutation data matrices from protein sequences, Comput. Appl. Biosci., 8 (1992) 275–282.
56. P.C. Ng, J.G. Henikoff, S. Henikoff, PHAT: a transmembrane-specific substitution matrix Predicted hydrophobic and transmembrane, Bioinformatics, 16 (2000) 760–766.
57. J. Overington, D. Donnelly, M.S. Johnson, A. Sali, T.L. Blundell, Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds, Protein Sci., 1 (1992) 216–226.
58. J.U. Bowie, R. Lüthy, D. Eisenberg, A method to identify protein sequences that fold into a known three-dimensional structure, Science, 253 (1991) 164–170.
59. P. Hogeweg, B. Hesper, The alignment of sets of sequences and the construction of phyletic trees: an integrated method, J. Mol. Evol., 20 (1984) 175–186.
60. G. Vriend, WHAT IF: a molecular modeling and drug design program, J Mol Graph, 8 (1990) 52–56, 29.

61. C. Chothia, A.M. Lesk, The relation between the divergence of sequence and structure in proteins, EMBO J., 5 (1986) 823–826.
62. C. Sander, R. Schneider, Database of homology-derived protein structures and the structural meaning of sequence alignment, Proteins, 9 (1991) 56–68.
63. G. Vriend, C. Sander, Detection of common three-dimensional substructures in proteins, Proteins, 11 (1991) 52–58.
64. Structural alignment - Wikipedia, the free encyclopedia, (n.d.).
65. P. Koehl, Protein structure similarities, Current Opinion in Structural Biology, 11 (2001) 348–353.
66. R. Kolodny, P. Koehl, M. Levitt, Comprehensive Evaluation of Protein Structure Alignment Methods: Scoring by Geometric Measures, Journal of Molecular Biology, 346 (2005) 1173–1188.
67. A.S. Konagurthu, J.C. Whisstock, P.J. Stuckey, A.M. Lesk, MUSTANG: a multiple structural alignment algorithm, Proteins, 64 (2006) 559–574.
68. M.E. Ochagavía, S. Wodak, Progressive combinatorial algorithm for multiple structural alignments: application to distantly related proteins, Proteins, 55 (2004) 436–454.
69. N.S. Boutonnet, M.J. Rooman, M.E. Ochagavia, J. Richelle, S.J. Wodak, Optimal protein structure alignments by multiple linkage clustering: application to distantly related proteins, Protein Eng., 8 (1995) 647–662.
70. N. Leibowitz, R. Nussinov, H.J. Wolfson, MUSTA--a general, efficient, automated method for multiple structure alignment and detection of common motifs: application to proteins, J. Comput. Biol., 8 (2001) 93–121.
71. M. Shatsky, R. Nussinov, H.J. Wolfson, MultiProt — A Multiple Protein Structural Alignment Algorithm, in: R. Guigó, D. Gusfield (Eds.), Algorithms in Bioinformatics, Springer Berlin Heidelberg, Berlin, Heidelberg, n.d. pp. 235–250.
72. O. Dror, H. Benyamini, R. Nussinov, H. Wolfson, MASS: multiple structural alignment by secondary structures, Bioinformatics, 19 Suppl 1 (2003) i95–104.
73. Y. Ye, A. Godzik, Multiple flexible structure alignment using partial order graphs, Bioinformatics, 21 (2005) 2362–2369.
74. C. Guda, E.D. Scheeff, P.E. Bourne, I.N. Shindyalov, A new algorithm for the alignment of multiple protein structures using Monte Carlo optimization, Pac Symp Biocomput, (2001) 275–286.
75. A. Andreeva, D. Howorth, J.-M. Chandonia, S.E. Brenner, T.J.P. Hubbard, C. Chothia, A.G. Murzin, Data growth and its impact on the SCOP database: new developments, Nucleic Acids Research, 36 (2007) D419–D425.
76. A.L. Cuff, I. Sillitoe, T. Lewis, A.B. Clegg, R. Rentzsch, N. Furnham, M. Pellegrini-Calace, D. Jones, J. Thornton, C.A. Orengo, Extending CATH: increasing coverage of the protein structure universe and linking structure with function, Nucleic Acids Res., 39 (2011) D420–426.
77. L. Holm, C. Sander, Dali/FSSP classification of three-dimensional protein folds, Nucleic Acids Res., 25 (1997) 231–234.
78. J. Parsch, J.M. Braverman, W. Stephan, Comparative sequence analysis and patterns of covariation in RNA secondary structures, Genetics, 154 (2000) 909–921.
79. U. Göbel, C. Sander, R. Schneider, A. Valencia, Correlated mutations and residue contacts in proteins, Proteins, 18 (1994) 309–317.
80. D.S. Marks, L.J. Colwell, R. Sheridan, T.A. Hopf, A. Pagnani, R. Zecchina, C. Sander, Protein 3D structure computed from evolutionary sequence variation, PLoS ONE, 6 (2011) e28766.
81. S.W. Lockless, R. Ranganathan, Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families, Science, 286 (1999) 295–299.

82. N.D. Clarke, Covariation of residues in the homeodomain sequence family, Protein Sci., 4 (1995) 2269–2278.
83. W.R. Atchley, K.R. Wollenberg, W.M. Fitch, W. Terhalle, A.W. Dress, Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis, Mol. Biol. Evol., 17 (2000) 164–178.
84. J.P. Dekker, A. Fodor, R.W. Aldrich, G. Yellen, A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments, Bioinformatics, 20 (2004) 1565–1572.
85. M. Olivier, R. Eeles, M. Hollstein, M.A. Khan, C.C. Harris, P. Hainaut, The IARC TP53 database: new online mutation analysis and recommendations to users, Hum. Mutat, 19 (2002) 607–614.
86. C. Béroud, T. Soussi, The UMD-p53 database: new mutations and analysis tools, Hum. Mutat, 21 (2003) 176–181.
87. C. Szabo, A. Masiello, J.F. Ryan, L.C. Brody, The breast cancer information core: database design, structure, and scope, Hum. Mutat., 16 (2000) 123–131.
88. P.D. Stenson, E.V. Ball, K. Howells, A.D. Phillips, M. Mort, D.N. Cooper, The Human Gene Mutation Database: providing a comprehensive central mutation database for molecular diagnostics and personalized genomics, Hum. Genomics, 4 (2009) 69–72.
89. T. Kawabata, The Protein Mutant Database, Nucleic Acids Research, 27 (1999) 355–357.
90. J.T. Eppig, J.A. Blake, C.J. Bult, J.A. Kadin, J.E. Richardson, The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse, Nucleic Acids Research, 40 (2012) D881–D886.
91. C.M. Eng, R.J. Desnick, Molecular basis of Fabry disease: mutations and polymorphisms in the human alpha-galactosidase A gene, Hum. Mutat., 3 (1994) 103–111.
92. G. Casey, The BRCA1 and BRCA2 breast cancer genes, Curr Opin Oncol, 9 (1997) 88–93.
93. D.P. Guimaraes, P. Hainaut, TP53: a key gene in human cancer, Biochimie, 84 (2002) 83–93.
94. E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, et al., Initial sequencing and analysis of the human genome, Nature, 409 (2001) 860–921.
95. J.C. Venter, The Sequence of the Human Genome, Science, 291 (2001) 1304–1351.
96. R. Drmanac, The advent of personal genome sequencing, Genet. Med., 13 (2011) 188–190.
97. J.E. Lunshof, J. Bobe, J. Aach, M. Angrist, J.V. Thakuria, D.B. Vorhaus, M.R. Hoehe, G.M. Church, Personal genomes in progress: from the human genome project to the personal genome project, Dialogues Clin Neurosci, 12 (2010) 47–60.
98. C. Alkan, B.P. Coe, E.E. Eichler, Genome structural variation discovery and genotyping, Nat. Rev. Genet., 12 (2011) 363–376.
99. What is the human variome project?, Nat. Genet., 39 (2007) 423.
100. S.-M. Ahn, T.-H. Kim, S. Lee, D. Kim, H. Ghang, D.-S. Kim, B.-C. Kim, S.-Y. Kim, W.-Y. Kim, C. Kim, D. Park, Y.S. Lee, S. Kim, R. Reja, S. Jho, C.G. Kim, J.-Y. Cha, K.-H. Kim, B. Lee, J. Bhak, et al., The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group, Genome Res, 19 (2009) 1622–1629.
101. R.K. Kuipers, H.-J. Joosten, W.J.H. van Berkel, N.G.H. Leferink, E. Rooijen, E. Ittmann, F. van Zimmeren, H. Jochens, U. Bornscheuer, G. Vriend, V.A.P.M. dos Santos, P.J. Schaap, 3DM: systematic analysis of heterogeneous superfamily data to discover protein functionalities, Proteins, 78 (2010) 2101–2113.
102. H.-J. Joosten, Y. Han, W. Niu, J. Vervoort, D. Dunaway-Mariano, P.J. Schaap, Identification of fungal oxaloacetate hydrolyase within the isocitrate lyase/PEP mutase enzyme superfamily using a sequence marker-based method, Proteins, 70 (2008) 157–166.

103. R. Kourist, H. Jochens, S. Bartsch, R. Kuipers, S.K. Padhi, M. Gall, D. Böttcher, H.-J. Joosten, U.T. Bornscheuer, The alpha/beta-hydrolase fold 3DM database (ABHDB) as a tool for protein engineering, Chembiochem, 11 (2010) 1635–1643.
104. A. Cerdobbel, K. De Winter, D. Aerts, R. Kuipers, H.-J. Joosten, W. Soetaert, T. Desmet, Increasing the thermostability of sucrose phosphorylase by a combination of sequence- and structure-based mutagenesis, Protein Eng. Des. Sel., 24 (2011) 829–834.
105. H. Jochens, D. Aerts, U.T. Bornscheuer, Thermostabilization of an esterase by alignment-guided focussed directed evolution, Protein Eng. Des. Sel., 23 (2010) 903–909.
106. H. Jochens, U.T. Bornscheuer, Natural diversity to guide focused directed evolution, Chembiochem, 11 (2010) 1861–1866.
107. D. Hasenpusch, U.T. Bornscheuer, W. Langel, Simulation on the structure of pig liver esterase, J Mol Model, 17 (2011) 1493–1506.

Technical Background

2.1 Introduction

The 3DM platform is a bioinformatics toolbox that can be used to create Molecular Class Specific Information Systems (MCSIS) for protein superfamilies. This chapter aims to provide an overview of the technical background of the 3DM platform, the steps required to create a new MCSIS with 3DM and describes the workflow of the main modules used for superpositioning of structures, correlated mutation analyses, mutation retrieval, and variant pathogenicity predictions.

The 3DM platform consists of three main parts; the backend that creates the alignments and loads the database, the database itself and the frontend that deals with all user interactions (fig. 2.1). Several external tools are integrated into the 3DM workflow such as WHAT IF for structural superpositions, determination of contacts and solvent accessibility, sequence alignments, and handling of PDB files; YASARA for the creation of structural representations; ClustalW for phylogenetic trees; MUSCLE for sequence alignments; and ImageMagick for image handling.

The 3DM platform and its dependencies consist of millions of lines of code and offer lots of functionality. A complete overview of all these functions is omitted for the sake of brevity, however a technical oriented in-depth overview of the main components of the 3DM platform is essential for the interpretation of 3DM analyses. This chapter aims to provide such a technical background perspective on several major components of the 3DM platform.

2.2 3DM Platform

Central within each 3DM system is a database that is used to store available superfamily data. The first section of this chapter therefore describes the characteristics of the 3DM databases and some background information on data stored in these databases. Section 2.3 describes, in detail, the process of creating a 3DM system for a superfamily. A twelve step process is outlined describing the initial setup, creation of the superposition, subfamily alignments, and data incorporation. These first two sections describe the basis of every 3DM systems and can be used to study the inner workings of 3DM more closely. Subsequent sections describe components added to the 3DM platform to perform various analyses. Section 2.5 describes the algorithm used for correlated mutation analyses and the visualization techniques used to present this data to the user. This section complements the applied research described in chapters 3 and 5 that use correlated mutation analyses to study the characteristics of proteins in various superfamilies. Mutation retrieval is the subject of section 2.6. A list of mutation retrieval tools are reviewed in addition to the steps discerned in the process. The final section contains more details on the DNA diagnostics parts of 3DM with an overview of the data types used by Validator in the 3DM analyses, a list of databases containing useful data and an overview of available tools. The Mutator and Validator modules described in sections 2.6 and 2.7 were used extensively in the project described in chapter 6.

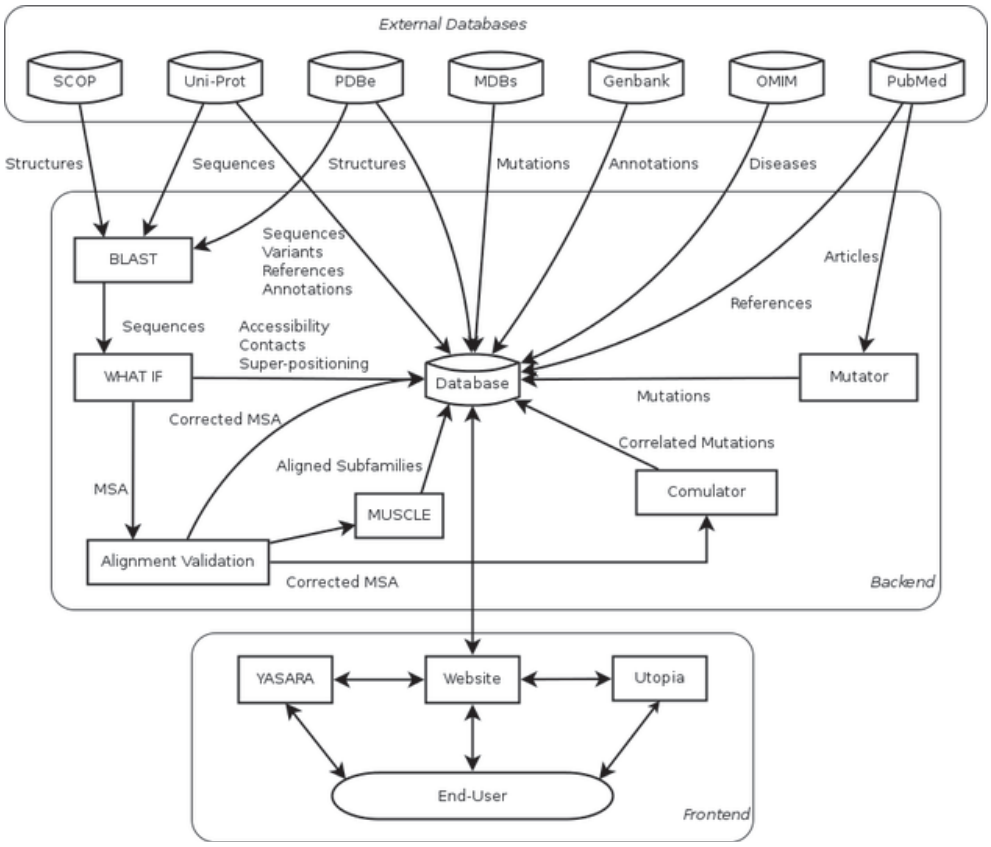


Figure 2.1: Overview of domains, modules and dataflows in the 3DM platform. Databases, both the 3DM database and external databases, are shown as cylinders, programs and scripts are shown in rectangles. Arrows indicate the flow of data between the modules. Bi-directional arrows indicate that data moves both ways. The separate domains within the 3DM platform are grouped in rounded squares. For more information on the 3DM database see chapter 2.3, the frontend and backend modules are described extensively in chapter 2.4. Usage of external databases is summarized in table 2.1.

3DM systems revolve around a structure-based multiple sequence alignment (MSA) that is generated specifically for each protein superfamily. The MSA, sequences, structures, annotations, contacts, mutations, and many other datatypes are gathered by the backend and stored in the database. Extensive cross-references ensure adequate interoperability of these data types. The generation of an MSA is a twelve step process that will be described in detail in chapter 2.4. In short, the generation of an MSA starts with obtaining all structures and sequences of superfamily members. All structures are superposed pairwise on a master structure using WHAT IF (fig. 1.7 a1) [1]. The superpositioning is used to determine the common structural core (fig. 1.7 a2), and to select templates for subsequent subfamily alignments (fig. 1.7 a3). Sequence alignments are generated for each subfamily to

incorporate closely related sequences for which no structure data are available (fig. 1.7 b). A complete superfamily alignment is assembled using the common core positions of all aligned sequences (fig. 1.7 c). External data from various databases and providers are incorporated into the database and linked to proteins, positions, nucleotides, or amino acids. All these external data are interconnected using the MSA as a frame of reference. Users can interact with the system using a web interface or suitable external viewers such as the YASARA structure viewer [2] or the Utopia PDF reader [3].

External Database	*Data Provided*	*Section*
SCOP	Classification of structures into superfamilies	2.4.2
Swiss-Prot/TrEMBL	Sequences, annotations, variants and cross-references	2.4.11
PDBe	Protein structures	2.4.2
Mutation Databases (MDBs)	Mutations & phenotypes	2.4.11
Genbank	Sequences, annotations, variants and cross-references	2.4.11
OMIM	Diseases	2.4.11
PubMed	Articles & references	2.4.11 & 2.6

Table 2.1: External databases used in the 3DM platform. For each database the primary use case is listed together with the section of this chapter where more information can be found.

Generation of 3DM systems should be as automated as possible to reduce the required domain knowledge, manual supervision by human experts, and total time needed for the generation of new 3DM systems. This has nearly been achieved for the generation of the MSA itself and for the incorporation of external data into the system, both of which are >99% automated. Generating a superfamily structure superposition however has not yet been fully automated. Determining the optimal superposition of two structures is a complex process, and small changes in the transformation often reduce the RMSD in one region while increasing it in another. It is therefore impossible to obtain a definitive superposition as the outcome will differ depending on the order in which the structures where compared. User interaction is especially required in the case of multi domain structures that can move relative to each other. Consequently, the determination of the common structural core is also non-trivial as the common core and the superpositioning are directly linked. For an alignment position to be incorporated into the common core, at least 80% of all structures are required to have a residue within 2.5A of the mean position. Any changes in the superpositioning therefore have a direct effect on which positions can be included in the common core.

2.3 3DM Database

Figure 2.2 shows a condensed overview of the 3DM database schema. As of spring 2012, the full schema contains 43 tables, 57 foreign keys, and 37 indices. The full database schema can therefore only be viewed interactively. 3DM systems focus on the proteins in one superfamily. The sequence of each protein is stored in the Protein table. Information for one protein is often scattered over many database so that multiple protein identifiers can be linked to a single sequence. Crambin, for example, is available as P01542 and CRAM_CRAAB in UniProt; 3NIR, 1AB1, 1CRN and others in the PDB; A01805 and KECX in PIR; 6226577 in the NCBI Protein database; UPI0000110BEA in UniParc; 0710210A in PRF; etc. Different external databases offer different annotations and different levels of details in their descriptions, or focus on different biological aspects. Annotations such as species, gene name, database of origin, and biological function are therefore stored separately for each protein identifier. As the different identifiers describe the same sequence, the annotations from the databases are often identical or highly similar. The annotation data is however, used for data analysis and clustering. The redundancy resulting from storing identical annotations multiple times is therefore outweighed by the ability to store a complete overview of annotations for each protein.

Each amino acid of each protein is stored in a separate record in the AminoAcid table. With an average protein length of 400 amino acids this results in, on average, 400 additional records for each protein. With superfamilies ranging in size from a few hundred,

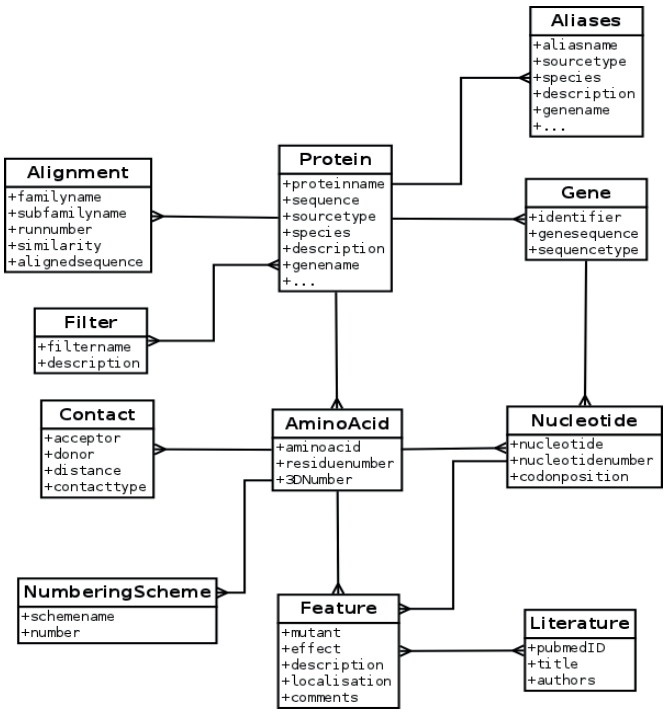


Figure 2.2: Condensed Enhanced Entity Relationship (EER) diagram of the 3DM database. Each block represents a table, with the name in bold at the top and the fields listed underneath. Lines indicate foreign key relations between the tables, with the line endings indicating whether the relationship is a one-to-one, one-to-many, or many-to-many key.

to tens of thousands of proteins this results in a large number of extra records. The cost of storing the amino acids separately is however clearly outweighed by the flexibility of being able to link many annotations and multiple numbering schemes to individual amino acids.

Amino acid annotations, descriptions and variant data such as mutations, conflicts, trans-membrane regions, active sites, glycosylation sites, phosphorylation sites, repeats, zinc-finger domains, etc. are stored in the Feature table. Each annotation is linked to the database from which it was retrieved and described using metadata. References to publications that are linked to the annotation are stored in the Literature table.

Figure 2.3: Numbering an aligned sequence. Sequence shown in aligned 3DM format with structurally conserved residues capitalized and in bold, and residues in structurally non-conserved regions in lowercase. Residue numbering is shown above, and alignment numbering below the sequence. Core and non-core (variable) regions are separated from each other by a space. Structurally conserved residues are numbered according to their position in the common core of the superfamily. As not all structures in the superfamily contain heterogeneous residues for all positions in the common core some positions can remain empty. In the example sequence, no heterogeneous residues are present for positions 34 through 37 of the common core. Variable regions are aligned and numbered individually for each subfamily. The numbering scheme for variable residues includes a character prefix based on the location of the region in the subfamily alignment. The first residue in the first variable region is numbered a1, the second a2, the first residue in the second variable region is numbered b1, etc. Variable regions often differ markedly in their makeup between aligned proteins. A loop between two structurally conserved areas, for example, might consist of up to 20 residues in one protein while another protein has no residues at all between the two areas. Variable regions lacking residues are left empty, marked with a minus, and skipped in the numbering such as region c in the example sequence. The fourth variable region in the example sequence is therefore prefixed with a d as the character prefix indicates the location of the variable region in the subfamily alignment rather than the location in the sequence. The process of assigning numbers to residues is completely automated.

Each amino acid is numbered using two different numbering schemes: its sequential residue numbering and the 3DM alignment position numbering. Sequential residue numbering starts at one for the first residue in the sequence and increases by one for each amino acid in the protein, independent of any alignment. Alignment position numbers depend on the location of the residue in the superfamily superposition with a sequential number assigned to each position in the common core. Residues in structurally variable regions are numbered separately with a region-dependent character prefixed to each residue (fig. 2.3). The alignment position number of a residue therefore indicates the position in the structure alignment rather than the position in the sequence of the protein. Residues in different proteins with the same alignment position number are structurally at a corresponding position. For core regions this comparison holds true for essentially the entire superfamily, variable positions however can only be confidently aligned within subfamilies. Custom numbering schemes for the alignment positions can be defined by the user. These schemes do not have to be ascending or numeric, and may instead consist of identifiers for structural elements such as helix1.1.

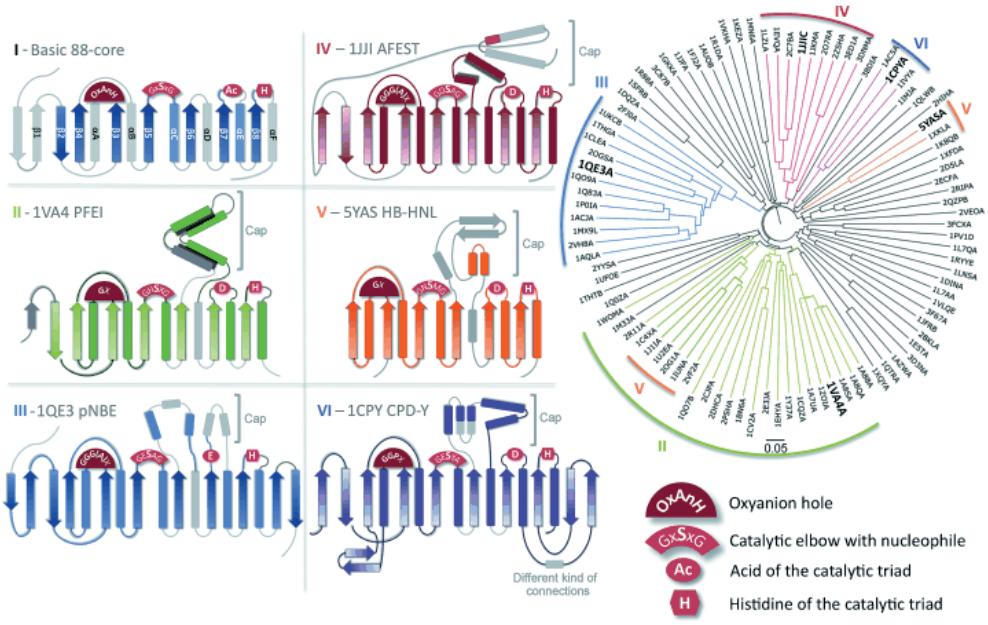


Figure 2.4: Families and folds in the alpha-beta hydrolase superfamily. 3DM superfamilies are divided into families based on structural folds that are further subdivided into subfamilies. A family represents a subset of the complete superfamily with a distinct structural component that is not conserved in the complete superfamily. The common core of family I consists of 9 beta-strands and contains 88 positions. This common core is shared by all superfamily structures. Families II-VI include additional positions from the beta-sheet and connecting sequences on top of the 88 positions of family I. The regions added to the other families represent the distinct folds of the families. The phylogenetic tree shown on the right contains all subfamily templates of the Basic 88-core family. Each template structure that is also a template in one of the other families is highlighted with the family's color.

Contact data such as disulfide bridges, dimer contacts, ligand contacts, and hydrogen bonds are obtained using WHAT IF as described in chapter 2.4.11. Contacts can be stored between two amino acids, or between an amino acid and a compound found in a PDB file such as a ligand or metal atom.

3DM systems focus on protein data but several types of genomic data are also included in the database. Genome sequences are stored in the Gene table with the individual nucleotides stored in the Nucleotide table. Nucleotides are stored separately for the same reasons as storing amino acids separately. Only a limited number of genes are stored because storing each nucleotide of each amino acid of each protein would result in an enormous amount of data, and genomic data is used only sparsely within 3DM. Genomic data is mainly used by Mutator to ground DNA mutations to genes. Published DNA mutations that are relevant to our users tend to originate from patient data, or as the result of testing drugs, treatments, or conditions on model animals. Therefore only genes from human and model animals such as rat and chimpanzee are stored.

Superfamilies in 3DM systems are divided into families based on structural folds (fig. 2.4). Families in turn are subdivided into subfamilies based on structural templates. Proteins can be aligned against multiple subfamilies due to similarities between the subfamily templates. The Alignment table contains protein alignments combined with alignment derived data such as the family, subfamily, and profile similarity for each alignment of each protein. Due to sequence similarities between subfamily templates many proteins can be aligned to multiple subfamily templates. All possible alignments are stored for each protein although only one alignment is used within the 3DM systems (see also chapter 2.4.9). Depending on the number of subfamilies and the sequence similarity between the subfamily templates a protein can therefore occur multiple times in the Alignment table.

Superfamilies can consist of many proteins. In practice, 3DM systems have currently been generated for superfamilies ranging from 100 to 115.000 proteins. Large superfamilies often consist of multiple sub-groups, often with different functions. The actin-like ATPase domain 3DM for example, contains several sub-groups including the hexokinases and glucokinases. It is not always desirable or practical to use all proteins when running an analysis or when visualizing data. The use of a subset of proteins can often even be beneficial as it allows the user to investigate and compare the differences between arbitrarily defined subgroups of the superfamily. A filter mechanism has been incorporated into 3DM to facilitate the investigation of properties of subsets or differences between subsets of the superfamily. A filter consists of a list of proteins that can be created by the user either through a number of search options or through manual selection. To increase flexibility upon filter creation several options are available, for example, filters can be combined, and all proteins from a filter can be removed from another filter. The filtering mechanism allows for a dynamic view of subsets of the 3DM system, and has proven a very useful tool for the investigation of superfamily characteristics.

2.4 Superfamily Creation

The creation of a 3DM system for a superfamily includes a large number of steps, most of which must be executed in the right order. Table 2.2 lists all major steps and indicates in which section of this chapter details about that step can be found.

Chapter	*Step*
2.4.1	Create 3DM Environment
2.4.2	Retrieve Structure Data
2.4.3	Superfamily Superpositioning
2.4.4	Update Structure Files With Common Orientation
2.4.5	Determine Structural Conservation
2.4.6	Select Subfamily Templates
2.4.7	Expand Common Core
2.4.8	Subfamily Alignments
2.4.9	Alignment Quality Control
2.4.10	Assemble Superfamily Alignment
2.4.11	Integrate Heterogeneous Data Types
2.4.12	Visualize Superfamily Data

Table 2.2: Major steps in the creation of a 3DM system for a protein superfamily. Numbers refer to the chapters were the step is explained in detail.

2.4.1 Create 3DM Environment

Some preparations have to be made before a new 3DM system can be created. A database as described in chapter 2.3 is created to store superfamily related data. A directory hierarchy is also created to serve as backend storage for files required or generated during the creation of a new 3DM system. Macros to automate processes using WHAT IF or YASARA, output files from the alignment process, and logfiles of most of the alignment steps are stored in these directories. The contents of these directories are generally only used to debug problems that occurred during the generation of the 3DM system. Users of 3DM systems do not need to interact with any of the files in the backend directories.

2.4.2 Retrieve Structure Data

All structures belonging to the superfamily must be collected before a structure alignment can be generated. A BLAST search against the PDB database is used with default settings to retrieve the structure most similar to the protein of interest. When multiple high scoring matches are found a structure is selected based on resolution or elucidation method. This master structure is used as the template for the structure alignment. A complete set of superfamily structures is subsequently assembled using SCOP and BLAST. The SCOP database contains PDB files organized in a superfamily hierarchy categorizing structures into classes, folds, and superfamilies. Each SCOP superfamily is subdivided into families,

domains, and proteins. SCOP is curated manually which results in a backlog of recently published structures that have not yet been included into the SCOP hierarchy. The most recent release of SCOP is currently 1.75, released in June 2009. We are therefore working on including family classifications from CATH [4] and perhaps the new HSSP database [5] to augment the SCOP classification.

The SCOP superfamily corresponding to the master structure is determined using the structure retrieved from the initial BLAST search against the PDB database. Additional BLAST searches are performed using the SCOP entries as queries to retrieve PDBs not included in SCOP. PDB files for all selected structures are retrieved from a local mirror of the PDB archive. FASTA files are generated for every chain of every structure for use in the alignment procedure.

2.4.3 Superfamily Superpositioning

All superfamily structures are superposed onto the master structure. The superpositioning is done using WHAT IF's MOTIF option which can be found in the SUPPOS menu [6] (fig. 1.7, a1) using default settings. The superpositioning procedure used by WHAT IF is explained in detail in chapter 1.2. The outcome of the superpositioning procedure is a

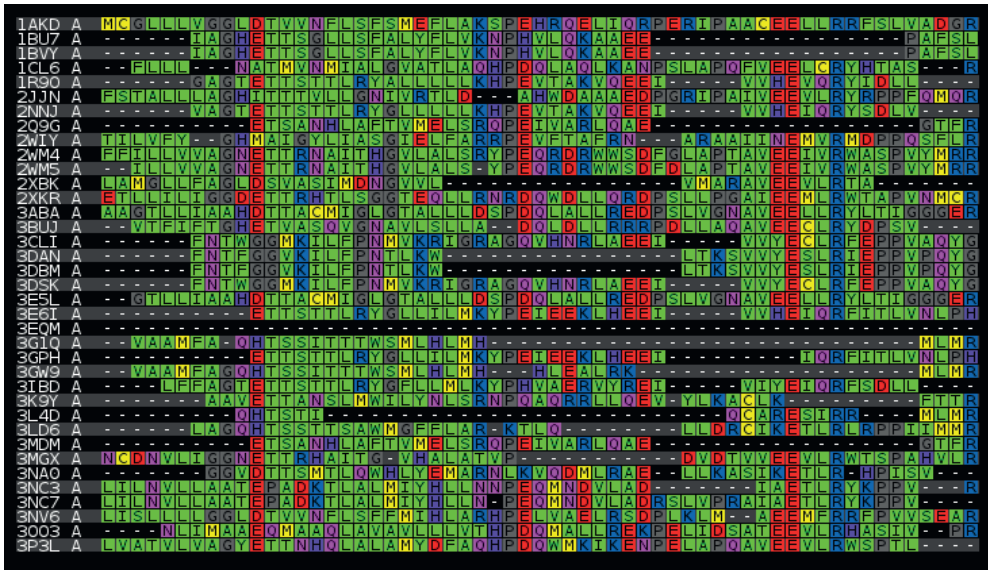


Figure 2.5: Structure-based sequence alignment for P450 superfamily members. The left column contains the PDB and chain identifiers of the structures. The first line contains the sequence of the master template. Subsequent lines indicate the sequence alignment that was obtained from the structure alignment. Positions in the template where the other structures do not have a homologous residue are marked with a minus. Amino acids are colored based on chemical properties.

WHAT IF database with all structures superimposed on the master template. This database is used to determine the common structural core and generated updated structure files.

2.4.4 Update Structure Files with Common Orientation

The orientation of structures in the PDB depends on several factors such as the elucidation method, experiment, and experimentalist. The origin of superfamily structures is very diverse and therefore these structures do not use a common orientation. A unified orientation of these structures allows structural scenes to be generated containing an arbitrary selection of superfamily structures (fig. 1.2 & 1.6). These scenes can, for example, be used to compare ligand docking between structures or study structural diversity by displaying several structures at once. WHAT IF is used to apply the transformations obtained from the superfamily superposition procedure to generate new structures with a common orientation. These new structures can then be used by other programs to display a common view of the superfamily.

All components from each PDB, such as protein chains, ligands and metals, are additionally stored in separate PDB files. For example, structure 1BSX will be stored in five different files: 1) the complete structure including protein and ligand chains, 2) two separate protein chain files, one containing protein chain A and one containing chain B, 3) two separate ligand files, one containing ligand chain A and one containing chain B. The newly reoriented structures are used as basis for the split PDB files which therefore also use the common orientation.

2.4.5 Determine Structural Conservation

The amount of structural conservation differs between regions of the superfamily structures. To determine the common structural core, all superfamily members are compared with the master template using WHAT IF (fig. 1.7, a2 & fig. 2.5). Structural conservation is defined as three consecutive residues in both structures for which the distance between the Cα atoms of each pair is less than 2.5Å. When three consecutive residues are found that match these criteria, WHAT IF checks if additional residues can be added without violating the RMSD constraints.

2.4.6 Select Subfamily Templates

Proteins for which no structure data are available are included into 3DM systems using sequence alignments. These sequences are grouped into subfamilies based on subfamily templates. Subfamily templates are selected based on sequence similarity. All superfamily PDB structures are examined in order. The first structure, usually the master template used for the superposition, is automatically selected as a subfamily template. For each subsequent structure the sequence similarity is determined with the list of selected subfamily templates. Structures are added as subfamily templates when they are less than 80% similar to all currently selected subfamily templates. The selection of subfamily templates therefore

depends largely on the order in which the structures are compared. A proper selection of subfamily templates probably requires a phylogenetic tree, evaluation of structure properties such as resolution, and in-depth knowledge of the superfamily. Smarter subfamily template selection will soon become necessary as more and more sequence and structure data become available and the list of potential subfamily templates continues to grow.

2.4.7 Expand Common Core

The initial common structural core is determined using a single superposition based on the master template. This common core is relatively conservative due to structural divergence and the flexibility of structural elements. Conformational changes in the loops connecting secondary structure elements, crystallization artefacts, or contaminations can shift structural elements beyond the 2.5Å cut-off used for the determination of common core residues. Additionally, hinges sometimes allow for much larger conformational changes without requiring large insertions or deletions of sequence. In the Nuclear Receptor family for example, helix 12 changes confirmation depending on activation or inhibition of the ligand binding domain. The composition of the helix itself is conserved, only the location in the structure is variable due to the hinge. Helix 12 can therefore be included in the common core based on sequence similarities despite the divergence in structural location. To expand the conserved regions determined in step 2.4.5 the common core is determined again using the subfamily templates structures selected previously.

SeqNo	ArbNo	A	OPEN	ELON	WEIGHT	V	L	I	M	F	W	Y	G	N	
1		N	0.20	0.40	100	-0.11	-0.11	-0.10	-0.06	-0.15	-0.15	-0.06	0.00	0.94	
2		L	0.20	0.40	100	0.09	0.94	0.10	0.14	0.10	-0.00	-0.01	-0.10	-0.11	
41		V	0.20	0.40	100	0.94	0.09	0.10	0.04	-0.00	-0.05	-0.01	-0.05	-0.11	
42		P	0.20	0.40	100	-0.06	-0.11	-0.10	-0.11	-0.10	-0.15	-0.16	0.00	-0.11	
43		D	8.00	1.00	100	-0.11	-0.11	-0.15	-0.06	-0.10	-0.15	-0.11	0.00	0.09	1
44		L	8.00	1.00	100	0.69	1.14	0.30	0.34	0.30	0.20	0.19	-0.10	-0.21	2
45		V	8.00	1.00	100	0.94	0.09	0.10	0.04	-0.00	-0.05	-0.01	-0.05	-0.11	3
46		W	8.00	1.00	100	-0.06	-0.01	-0.00	-0.11	0.15	1.00	0.14	-0.10	-0.16	4
47		T	8.00	1.00	100	-0.01	-0.01	-0.00	-0.01	-0.10	-0.05	-0.11	-0.05	-0.01	5
48		R	8.00	1.00	100	-0.06	-0.06	-0.10	-0.11	-0.10	-0.00	-0.06	0.00	-0.01	6
49		C	8.00	1.00	100	-0.11	-0.16	-0.10	-0.01	-0.15	-0.05	-0.11	-0.10	-0.11	7
50		N	0.20	0.40	100	-0.11	-0.11	-0.10	-0.06	-0.15	-0.15	-0.06	0.00	0.94	
51		G	8.00	1.00	100	-0.06	-0.11	-0.10	-0.11	-0.15	-0.10	-0.16	0.96	-0.01	11
52		G	8.00	1.00	100	-0.06	-0.11	-0.10	-0.11	-0.15	-0.10	-0.16	0.96	-0.01	12

Figure 2.6: Partial alignment profile. Columns show the residue number (position 3-40 not shown), residue, gap open penalty, gap elongation penalty, position weight, score for each amino acid (some scores omitted), and alignment position numbers. The gap open and elongation scores are adjusted for positions 43-49, and 51-52, as these positions are part of the common core. These regions are used as anchors in the sequence alignment and therefore the scores have been adjusted to prefer insertions and deletions in variable regions.

2.4.8 Subfamily Alignments

Subfamily alignments are created for each of the templates selected in step 2.4.6 using a profile based iterative method (fig. 1.7b). These alignments are used to align proteins to the

closest related template structure in the superfamily. BLAST searches against the UniProt and PDB databases are used to preselect sequences which can potentially be aligned to the subfamily template. A subset of sequences is used for the alignment procedure instead of all sequences from PDB and UniProt to speed up the alignment procedure. The alignment procedure is profile based and iterates through multiple rounds. Using the template sequence an initial profile is created (fig. 2.6) for the first alignment round. Profiles for the 2nd, 3rd, and 4th rounds are based on the results of the previous iteration. After the first alignment is complete, a profile is created for the 2nd round using all sequences which were aligned to the profile in the first round with at least 70% identity. Profiles for the 3rd and 4th round are created using sequences 55% and 45% similar to the profile of the previous round. Several modifications are made to the profiles to improve the alignments. Common core regions in the template are used as anchors in the alignments. Insertions and deletions in

Alternative Alignments

Class	Round	Similarity	1	2	3	4
2FH8A	4	0.47	161 SAAGKNF	39 GGDIEGLIKQL	0 DYIKN	1 GTTALWLTP
1GVIA	4	0.43	161 SAAGKNF	39 GGDIEGLIKQL	0 DYIKN	1 GTTALWLTP
2AAAA	4	0.43	165 KNFYFVL	35 GGDIEGLIKQL	0 DYIKN	1 GTTALWLTP
1A47A	4	0.42	165 KNFYFVL	35 GGDIEGLIKQL	0 DYIKN	1 GTTALWLTP
2DH3A	4	0.41	3 SALPRYF	8 AISLVGVGLHA	189 DYIKN	1 GTTALWLTP
1LWJA	4	0.40	152 DIVGAPF	48 GGDIEGLIKQL	0 DYIKN	1 GTTALWLTP
1EA9C	4	0.40	161 SAAGKNF	39 GGDIEGLIKQL	0 DYIKN	1 GTTALWLTP
1WZAA	4	0.40	152 DIVGAPF	48 GGDIEGLIKQL	0 DYIKN	1 GTTALWLTP
1BVZA	4	0.40	161 SAAGKNF	39 GGDIEGLIKQL	0 DYIKN	1 GTTALWLTP
1H3GA	4	0.39	161 SAAGKNF	39 GGDIEGLIKQL	0 DYIKN	1 GTTALWLTP
1QHOA	4	0.39	161 SAAGKNF	39 GGDIEGLIKQL	0 DYIKN	1 GTTALWLTP
1IZJA	4	0.38	161 SAAGKNF	39 GGDIEGLIKQL	0 DYIKN	1 GTTALWLTP
1BF2A	4	0.38	191 DKNDHGF	9 GGDIEGLIKQL	0 DYIKN	1 GTTALWLTP
1UOKA	4	0.37	161 SAAGKNF	39 GGDIEGLIKQL	0 DYIKN	1 GTTALWLTP
1M53A	4	0.37	161 SAAGKNF	39 GGDIEGLIKQL	0 DYIKN	1 GTTALWLTP
1BLIA	4	0.37	169 FVLTDRF	8 RGGAKDGDKND	23 DYIKN	1 GTTALWLTP
1HVXA	4	0.37	169 FVLTDRF	8 RGGAKDGDKND	23 DYIKN	1 GTTALWLTP
1UD2A	4	0.36	169 FVLTDRF	8 RGGAKDGDKND	23 DYIKN	1 GTTALWLTP
2E8YA	4	0.36	1324 RMIYELH	20 AFTQSGSKSV-	0 MHLKE	4 GLNSVHLLP
1EH9A	4	0.36	191 DKNDHGF	9 GGDIEGLIKQL	0 DYIKN	1 GTTALWLTP
1W9XA	4	0.36	169 FVLTDRF	8 RGGAKDGDKND	23 DYIKN	1 GTTALWLTP

Figure 2.7: Alignment data of an alpha amylase protein. The superfamily consensus is shown at the top together with the subfamily consensus and template sequence of subfamily 2FH8A. Alignments for the protein against several subfamily templates are shown in the table. Each line contains the alignment of the protein against a subfamily profile. The alignments are ordered by profile iteration and similarity. Each line shows the subfamily, profile iteration, similarity and aligned common core regions of the protein. Common core regions are numbered at the top. The numbers between the common core regions indicate the number of variable residues. The selected alignment used in 3DM is shown in green. The other alignments are only used for quality control. Alignment positions that do not match the selected alignment are highlighted in red.

common core regions are less likely as these regions are structurally conserved. Therefore the alignments are tuned to favour insertions and deletions in variable regions by adjusting the gap open and gap elongation penalties. The iterative profile based alignment procedure is performed twice for each subfamily. The first run consists of three iterations and the second run of four. Superfamily statistics such as conservation and hydrophobicity are determined for all positions of the common core after the first three iterations. During the second set of alignments the superfamily statistics are used to tune the profiles. For example, a highly conserved serine at a structurally conserved position will get a positive bonus score and hydrophobic residues receive a small bonus score on predominantly hydrophobic positions. Due to differences in length of the sequences in the variable regions these regions cannot be properly aligned during the profile alignment phase. MUSCLE [7] is therefore used to align the sequences of each variable part of each subfamily separately using 4 iterations.

2.4.9 Alignment Quality Control

A sequence can be aligned to several templates in the superfamily due to sequence and structure similarities between templates (fig. 2.7). For each protein in the 3DM superfamily only one alignment is used. Proteins with ambiguous alignments are therefore checked to obtain the highest quality alignment for each protein. An alignment is initially selected for each protein based on profile iteration and similarity. In figure 2.7 for example, the alignment against profile iteration 4 of subfamily template 2FH8A is preferred over the other alignments as the similarity is higher. All proteins for which the selected alignment has a profile similarity over 50% are assumed to be correct. The alignment consistencies of proteins with a selected alignment with a similarity below 50% are checked and if necessary adjusted using the alternative alignments. If only one alignment is available for a protein and the similarity of that alignment with its profile is below 50% the protein is removed from the database as its alignment cannot be verified. The 50% similarity cut-off was chosen due to the high number of comparisons the alignment consistency check requires for each protein, and the low rate of inconsistently aligned sequences with similarities over 50%.

To check the quality of a proteins alignment the alterantive alignments for each region of the common core are compared with each other. A consistent alignment is not necessarily an indication of a high-quality alignment. Inconsistent alignments however, indicate regions for which multiple alignment solutions were available that scored evenly. The alignments of these regions are less certain then the unambiguously aligned regions and have to be checked to ensure the alignment is correct. To check the alignment consistency of a common core region the alternatives for each region are determined and evaluated using the subfamily and superfamily consensus sequences. A similarity score is determined for each of the common core regions with the subfamily template and superfamily consensus and used to selected the best alignment for each region. The final sequence for a protein can thus consist of common core regions from alignments against different subfamily templates. After assembly, the overall score of the alignment is calculated and used to determine if

the alignment passes quality control. Proteins for which the alignment does not score high enough on this final test are removed from the database.

The alignment of a common core region is considered correct and left unchanged when it is consistent between different subfamilies. In the example shown in figure 2.7 this is the case for common core regions 3 and 4 which are essentially aligned unambiguously. Common core regions that have been aligned inconsistently are evaluated using the alternative alignments (fig. 2.7, regions 1, 2 and 5). For each inconsistently aligned common core region, the alternative alignments are selected based on frequency of occurrence. Six possible alignments are available for region 1 in figure 2.7: SAAGKNF 9 times, FVLTDRF 4 times, DIVGAPF, DKNDHGF and KNFYFVL twice and SALPRYF once. Only SAAGKNF and FVLTDRF occur more than the occurrence cut-off of 3 and are therefore considered as potentially correct alignments for this common core region. Both alignments are compared with the subfamily template and superfamily consensus sequence for common core region 1 using a BLOSUM62 matrix. Additional points are awarded for alignments that result in the lengths of the variable regions upstream and downstream of the common core region to match the length of the corresponding variable regions in the subfamily template. In this case, the variable regions of the subfamily template sequence flanking common core region 1 are 262 and 22 residues in length. Neither of the two alternative alignments results in flanking variable regions of these lengths so no bonus points are added. The alignment with the highest score is used in the final alignment. For the example case the final score for SAAGKNF is -5 and the score for FVLTDRF is +5 meaning that the FVLTDRF alignment will be used for the first region of the common core in this alignment. After each core has been checked, and if necessary replaced, a final evaluation of the updated alignment is performed to check the quality of the complete alignment. The scores of the correct common core regions are checked and alignments for which more than 2 regions score negatively are removed.

2.4.10 Assemble Superfamily Alignment

To construct a superfamily alignment the subfamily alignments must be combined as described in figure 1.7. The common core regions of all subfamily templates are already aligned automatically because of the superfamily superpositioning. The common core of each superfamily is based on structure conservation resulting from evolutionary pressure. The common core residues are therefore present in most aligned proteins. Residues aligned to the common core positions in the subfamily templates are extracted for each protein and used to construct a core alignment. A numbering scheme based on these core regions is applied to facilitate the transfer of information between structurally equivalent residues.

2.4.11 Integrate Heterogeneous Data Types

A large number of different data types and datasets can be integrated in a 3DM system using the MSA as basis. This ranges from additional descriptions of the proteins in the superfamily

to correlations between different alignment positions. Data types that can be integrated into 3DM systems include: mutations and variants, protein and residue annotations, inter- and intra-molecular contacts, ligand-, ion-, and, metal-contacts, genomic sequences, charge clamps, hydrophobicity, accessibility, RMSD of positions in the common core, B-Factor, cross-references, etc.

1) Publications. Articles are used as reference in 3DM systems whenever possible. Annotations retrieved from external databases for example, are sometimes linked to articles. Mutator, the 3DM tool used to extract mutations from literature also provides mutations annotated with articles. These articles themselves are not stored in 3DM systems, however a reference is stored for each articles linked to an annotation or variant. For each article the title, authors, PubMed identifier, and reftag are stored. Mutator (chapter 2.6) and the GenBank and UniProt parsers insert the literature data in the database whenever a referenced annotation or variant is stored.

2) Structures. Structures are incorporated into the alignment and stored as proteins in the Protein table. PDB files, however, contain more than just proteins. Compounds commonly found in PDB files include ions, metals, solvents, water molecules, ligands, DNA/RNA strains, and drugs. The rules used to identify the various compounds are listed in table 2.3.

Compound	*Identification rule*
Protein	Compound name starts with 'Protein'
Water	Compound name contains 'water'
Oxygens	Compound name contains 'O2'
DNA/RNA	Compound name is 'DNA/RNA'
Ions	Predefined list of ions including 'BA', 'BR', 'CA', 'CD', etc.
Metals	Predefined list of metals including 'AU', 'CO', 'FE', 'HG', etc.
Solvent	Compound name is any of 'EOH', 'GOA', 'GOL', or 'SO4'
Ligand	Compound is composed of 7 or more HETATMs
Unknown	Everything else not matched by any of the other rules

Table 2.3: Identification rules for compounds in PDB files. Rules for the identification of compounds commonly found in PDB files in the order they are used to identify compounds.

All compounds from the PDB file are stored in the 3DM database except water molecules, solvents, and unknown compounds. Additionally, each residue of each protein chain is stored separately to be able to link residues in PDBs with amino acids in the alignment

and link structure specific data to the PDB residues. These structure specific data include 3D-coordinates, secondary structure, b-factor, RMSD, accessibility, and the number of the residue in the PDB file. The complete PDB file is analysed and inserted using a PHP script. Each structure is loaded into WHAT IF to determine the individual components using the SHOSOU option. A second script parses the split PDB files generated after the superpositioning and inserts each residue from the PDB file linked to an amino acid from the alignment. Extraction and determination of structure specific data is described separately.

3) Mutations and variants. Mutation data is available from several sources such as literature and mutation databases. Mutator was developed for the selection and retrieval of relevant articles, scanning for mutations and grounding these mutations to residues in the database (chapter 1.4 and 2.6). UniProt and GenBank records of superfamily proteins also often contain mutations and sequence variants. XML formatted files containing all data on each protein are available for all entries from these two databases. Mutations and variants are extracted from these files and stored in the database as a feature linked to the relevant amino acid. A PHP [8] script has been written that uses SimpleXML data structures and XPath [9] queries to retrieve the variants from the XML files. UniProt records are available in XML format using a webservice. GenBank records are retrieved using the EFetch facility of the Entrez suite [10]. Records from both UniProt and GenBank are cached for one month to prevent excessive network traffic.

An additional source of mutation and variant data are custom databases containing datasets for species or diseases. A growing number of databases containing natural variants are also available due to the rapidly decreasing costs of sequencers. Hospitals and universities are increasingly participating in genome, or exome sequencing projects to diagnose patients or to create natural variant databases. These databases are intended for SNP detection and variation studies, and are an important source to study the natural variations in the human genome. For some high-profile diseases, specific mutation databases are available such as the iARC [11] database that contains mutations related to cancer. Aggregate databases for entire species are also available such as the HGMD for human and the Sacharomyces Genome Database for yeast [12]. These custom mutation databases are available for a limited number of species, diseases, and superfamilies as assembling, curating, and maintaining these databases takes a lot of manual work. Free public access to these databases is therefore not always available, and access methods and data formats differ significantly due to a lack of common standards. An often used file format to export data from databases and spread sheets is the Comma-Separated Values (CSV) format [13]. A PHP script was written for 3DM to parse CSV files, convert the mutation data, and store the data in the 3DM database. The mutations are processed by providing the script the location where the data can be retrieved and a description of the format. Databases which do not offer CSV or XML output are scripted individually.

4) Protein and residue annotations. Many annotations are available from the UniProt and Genbank records discussed in the previous paragraph. These annotations describe residues, domains, and proteins and include post-translational modifications, species, gene names, synonyms, binding sites, active sites, NCBI taxonomy id, taxonomic lineage, trans-membrane regions, etc. Annotations are linked either to individual amino acids or to complete proteins depending on their type. The script described in the previous paragraph is also used to extract annotations from UniProt and GenBank records, link the annotations to a protein or amino acid, and store the annotations in the 3DM database. A separate PHP script was made to extract annotations such as species, secondary structure, or mutations from PDB structure files. In contrast to the scripts used to determine for example contacts from PDB files, the extraction of annotations from PDB records does not require the coordinates of all atoms. Partial XML formatted records which lack the coordinate section are therefore used rather than complete PDB files for extraction of PDB data. XML records are more consistently structured then regular PDB files, and it is therefore much easier to retrieve data from these records compared to PDB raw files. XML formatted records containing only the PDB header are available from PDBe [14].

5) Contacts. Many types of contacts are formed between different amino acids and between amino acids and other molecules. Example contacts are hydrogen bonds, ligand contacts, disulfide bridges, and intra-molecular contacts. WHAT IF is used to determine these contacts using the SHOHBO, SHOCYS, and CONTACT options [15]. A WHAT IF macro to determine contacts is generated by a PHP script. The macro instructs WHAT IF to load one or more structures one by one, determine the contacts, and store the results in a separate file for each structure. The files containing the contacts are subsequently parsed to extract the raw data. The PHP script matches the residues in the structure with the residues in the database and stores the contacts. Hydrogen bonds, intra-molecular contacts, and cysteine bridges are linked to both amino acids involved in the contact. Ligand contacts are linked to an amino acid and to a PDB compound. The compounds can be anything co-crystallized with the protein in the PDB file such as hemes, activators, drugs, etc. Contact data such as cysteine bridges and binding sites are also available from UniProt and GenBank protein records. The parser described in the mutations and variants paragraph of this chapter is also used to extract contacts from GenBank and UniProt and store them in the 3DM database.

6) Genomic data. The PICR webservice [16] hosted by the EBI is used to match proteins in the 3DM database to gene sequences from RefSeq. RefSeq entries include the full gene transcript that consists of the promoter, coding sequence, poly-A tail, and 3'-UTR region. Unfortunately, a single UniProt accession code may return multiple RefSeq entries due to isoforms. RefSeq stores each transcript separately, UniProt however combines all isoforms of a protein in one record. As the numbering of isoforms between the databases is not

consistent, an amino acid translation of the transcript sequence is used to determine the correct match. A custom PHP script is used to invoke the PICR webservice, determine the correct isoform, and insert the data into the 3DM database.

7) Correlated mutations. Residue positions that appear correlated in alignments are potentially functionally related. Comulator is used to determine which residue pairs are actually correlated. The evolutionary basis of correlated mutations is described in chapter 1.3 and the implementation of the algorithm used to determine correlations in chapter 2.5.

8) Charge clamps. Charge clamps are commonly determined by investigating charged residues in close proximity of each other in structures. However, since most sequences do not have structural data associated with them, and we require an automated method to determine charge clamps we decided to use alignments instead to predict possible charge clamps. Charge clamps are defined as two residues in close proximity which contain an opposite charge. We defined lysine and arginine as positively charged, aspartic acid and glutamic acid as negatively charged, and all other amino acids as neutral. A script was written which for each pair of alignment positions determines the residue pairs that occur. A score is calculated based on the categories the occurring residues fall into. Pairs of residues with opposite charge are indicative of a possible charge clamp and score positively. Positions which have combinations of a neutral and a charged residue, and combinations with identically charged residues are indicative of non-charge clamped pairs and score negatively. Pairs of neutral residues are used to normalize the score. As the common core residues in the sequence alignments are linked directly to the structural alignment, a future expansion of the charge clamp code will include the distance between the residues in the structures in the formula to attempt to validate predicted charge clamps.

9) Hydrophobicity. The average hydrophobicity of alignment positions can be calculated using the amino acids that occur at the position. Many different scales are available to determine the hydrophobicity of an alignment position aimed at different situations and purposes. We decided to use a simple generic scheme in which each amino acid is either hydrophobic or hydrophillic determined by the amino acid type. A hydrophobicity score is determined for each alignment position by adding a point for each hydrophobic residue (WFALYIVM) and subtracting a point for each hydrophilic residue. Finally, the score for each position is normalized to 100 using the number of sequences available at that position.

10) Accessibility. Solvent accessibility of each alignment position is determined using WHAT IF. A PHP script creates a WHAT IF macro that passes all structure files from the

superfamily to WHAT IF one after the other. WHAT IF determines the vacuum accessibility of each atom of each residue using the custom VAC3DM option. The results are stored in an easily usable format in a separate file for each structure. The PHP script parses all output files, determines the total accessibility in percentages of the side chain of each residue, and averages the side chain accessibility for all residues on an alignment position. Both the individual accessibility scores for each residue of each structure, and the average accessibility score of each alignment position are stored in the 3DM database.

11) RMSD. The RMSD of an alignment position can be used as an indication of its flexibility and freedom of movement. To determine the RMSD of an alignment position, the coordinates of the C-Alpha atoms from the structures are used. A PHP script parses the structure files for a superfamily and extracts the ATOM records for all CA atoms. The RMSD is determined by a second PHP script using the atom coordinates. The RMSD of each PDB residue is stored in the 3DM database. An averaged RMSD is also stored for each alignment position.

12) B-Factor. The B-Factor is stored in PDB files in a separate data column. A PHP script is used to extract the data for each residue of each PDB file and store this data in the database. An averaged B-Factor value for each alignment position is also determined and stored.

13) Cross-references. Databases containing data, references, and annotations for all sorts of biological entities are scatter over the internet. To include all of this data in 3DM is not practical due to size and time constraints. Most of the available data has only a very small range of applications, and storing all data would mean enormous datasets. 3DM therefore includes references to many external databases to offer as complete an overview and description of each protein and amino acid as possible. References to external databases can be retrieved using PICR, or from the UniProt, GenBank and PDB records. Example references included in 3DM are: InterPro, Gene3D, PDBSum, ProSite, and PFam. In addition to the identifier, a description of the remote record is also stored to facilitate searching and data clustering. The PHP scripts used to retrieve the references from the various sources are described in the previous paragraphs.

14) Accession status. UniProt protein records contain not only information on proteins but also include updates and deprecations of the record themselves. Proteins in UniProt are not static and may be replaced, or removed over time. Several possibilities lead to the removal of accession codes from UniProt, these are: the merger of separate isoform records into a single record, retraction of the underlying genomic dataset, and removal of predicted ORFs

which after curation are found not to encode for proteins. Retrieving a record for a UniProt accession code using the webservices therefore has three possible outcomes: the record itself is returned, the record has been deprecated and a different record into which the entry has been merged is returned, or the record has been removed entirely and an empty record is returned. To keep the 3DM systems as up-to-date as possible the churn of accession codes is handled automatically by the UniProt retrieval script. When a different record was returned then was requested, the original accession entry is not removed from the database directly. Instead, the record is changed into an alias of the UniProt accession code into which it has been merged. When the webservice returns an empty record a second record in RDF format is retrieved to check the records status. If an accession code has been removed from the UniProt database this is reflected in the RDF record with an "obsolete" tag. These proteins are removed from the 3DM database completely as they do no longer exist.

2.4.12 Visualize Superfamily Data

Data in 3DM systems can be visualized using either the interactive website or the YASARA structure viewer (fig. 2.1). The website allows users to view the alignment, alignment statistics, correlations, protein and amino acid annotations, phylogeny trees, mutations, contacts, taxonomy trees, etc. Various tools are also provided to search data, analyse mutants, create homology models, and view data in structures. Most data can be visualized and retrieved from multiple different starting points. For example, mutations are listed on amino acid, protein, and alignment position detail pages. Export options are also available for various data types to allow the user to process data in external programs. Filters can be used to create subsets of the superfamily alignment based on a variety of options. These subsets can be used to limit the data visualised on the website to allow for more in-depth study of parts of the superfamily.

Data from the 3DM system can be mapped onto structures using the view data in structure tool (fig. 2.8). An arbitrary selection can be made of structures and homology models and 3DM data can be mapped on the residues in the structures. The orientation of structures and models is identical due to the reorientation of structure files during the creating of the 3DM system. Using the alignment position numbering in the structures and models ensures residues can be compared with each other and with the data on the website easily. A 3DM plugin is available for YASARA which can retrieve data from the database and plot these data interactively in the scene. The python plugin communicates with a series of XML producing PHP scripts on the 3DM webserver to obtain the requested data. The plugin allows users to map alignment position data such as conservation and correlations onto the structures, colour the structures based on various properties, and add structures and drugs to the scene on the fly. Protein and amino acid annotation data can also be retrieved and visualized using the plugin.

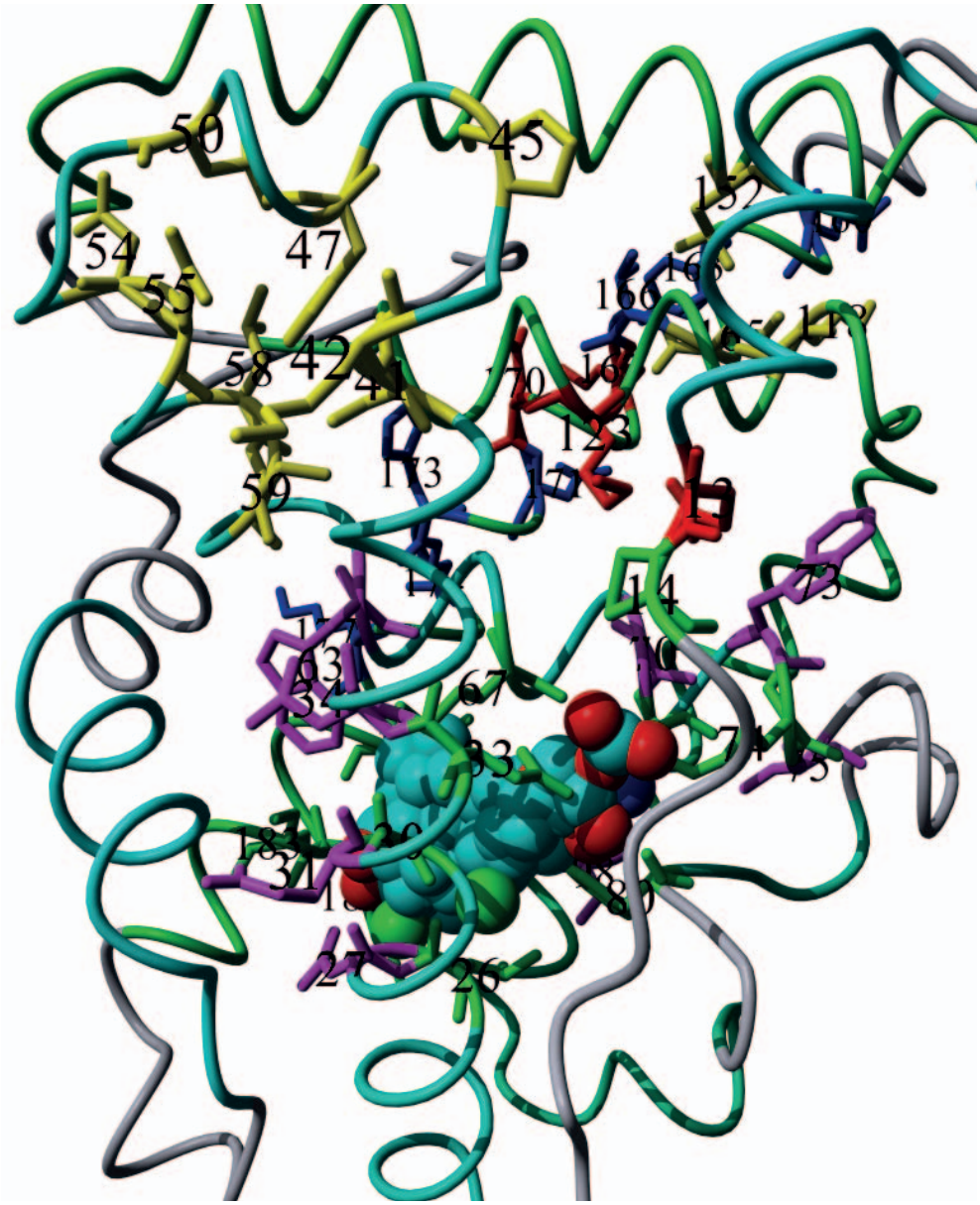


Figure 2.8: 3DM generated YASARA scene containing several examples of 3DM data plotted into structures. Scene contains Nuclear Receptor structure 1A28 chain A combined with ligands from 1BSX, 1DKF, 1ERE, and 1FBY superposed in the active site. Ligands are shown in atom sphere representation, residues in tube representation. Residues of interest have their side chains shown in atom sphere and are labeled with the alignment numbering. Common core regions are colored in a gradient from green to blue. Residues in variable regions are colored grey. Coloring of residue side chains based on plotted data: red) charge clamps, blue) intramolecular contacts, green) ligand contacts, yellow) conservation, magenta) correlated mutations.

2.5 Correlated Mutations

A Correlated Mutation Analysis (CMA) score is determined for each pair of alignment positions (formula 1). The CMA algorithm uses a double loop over the alignment positions to compare each position with all other positions. CMA scores are calculated bi-directionally and subsequently averaged for each pair of positions. The CMA scores are calculated using a custom developed C++ program (fig. 2.1, comulator). The CMA score is normalized by assigning the highest correlating pair a score of 1 and scaling the other scores correspondingly. Both the absolute and relative CMA scores are stored in the 3DM database. CMA scores are not comparable between different protein superfamilies as the outcome of the algorithm depends heavily on the number of sequences in the alignment.

$$CMA = \sum_{x=1}^{20} \sum_{y=1}^{20} |F| * N_{xy} \qquad with \qquad F = \frac{N_{xy}}{N_y} - \frac{N_x}{N_{(tot)}}$$

Formula 1: CMA scoring algorithm. The CMA score is determined for each pair of alignment positions. The absolute fraction difference (F) is multiplied by the observed frequency of the amino acid pair (N_{xy}) for each pair of amino acids at an alignment position. The results are summed to determine the CMA score for the position. The fraction differences are determined using the total number of sequences in the alignment (N_{tot}) and the total number of sequences that have amino acid x (N_x) or amino acid y (N_y) at that position.

CMA scores are visualized in the 3DM websites using heatmaps. A PHP script was written to generate the images using the precalculated CMA scores. Heatmaps containing all available datapoints for a superfamily are very large. The Nuclear Receptor superfamily for example, has 157 positions in the common structural core which results in a heatmap containing 24,649 datapoints (157 * 157). A condensed overview image of the correlations is therefore also included. This condensed overview contains only those positions for which any of the correlations exceeds the relative CMA score cut-off of 0.8. For the NR this results in 33 positions being included leading to a much more manageable heatmap consisting of 1,089 datapoints. Amino acid abundances can be visualized for each pair of correlating alignment positions. These abundances are displayed in a table with the amino acids of the two positions on the axes and the percentage of superfamily sequences that contains that amino acid pair in the cells.

2.6 Mutation Information

Mutator was developed to extract mutations from the literature. The aim was to automatically retrieve text from articles that describes possible amino acid or nucleotide mutations and to link these mutants to the correct proteins. Mutator consists of a series of python scripts designed to retrieve scientific articles from PubMed, to convert the articles to a computer-readable format, to extract the mutations, to ground the mutations to proteins (grounding

is the process that associates a mutation with the corresponding gene), to validate the mutations and discard false-positives, and to store the mutations in the 3DM database. Therefore the design criteria for Mutator are strongly influenced by the requirements of the 3DM platform. The main criteria for the design of Mutator were: applicability to complete superfamilies consisting of thousands of proteins, applicability to all superfamilies without requiring superfamily specific logic, and full-text article support. Mutator additionally has to be over 99% automated to be able to keep up with the rate of publication of new articles. At the time work on Mutator began in 2006, a number of tools where already available to deal with one or more parts of the mutation retrieval process (table 2.2). An overview of these tools, and additional tools available today is presented in the next section. The strong and weak points of the various tools will be listed together with the lessons we took from them.

2.6.1 Tools

Several tools have been developed to automate one or more steps of the mutation retrieval process (table 2.2). These tools work on different databases, use different technologies, and different levels of automation and user interaction. MuteXt [17] uses a computational method that identifies and extracts mutation data from scientific literature related to two specific protein families. MEMA [18] was developed to extract mutations and gene names from MEDLINE. Yip *et al.* [19] describe a method to update Swiss-Prot variants with new literature references. CoagMDB [20] is a database of mutations for five serine proteases involved in blood coagulation. MutationGrab [21] explores the use of network graphs to ground ambiguous mutations and MuGeX [22] retrieves mutations from MEDLINE that are related to a specific disease. Vtag [23] and Mtag [24] are applications for finding and highlighting biomedical entities in a text. Vtag looks for mutations and Mtag for cancer related objects and concepts.

Each of these tools has been developed with a specific goal in mind. Some tools focus on just one aspect of the mutation extraction process, while others are targeted at specific proteins, or showcase a novel technique for one of the steps of the mutation retrieval process. Most tools are able to retrieve articles, extract mutations, and ground the mutations to a protein. MEMA and MuteXt for example, offer a complete workflow that ranges from searching and retrieving articles to mutation extraction and validation. Vtag and Mtag instead focus on highlighting mutations in articles for manual extraction. The mutation source of the tools also varies, some use MEDLINE and others use PubMed. Most tools support data extraction from abstracts while some also have additional support for extracting data from full-text articles. The target proteins for which the applications were designed differ widely. CoagMDB and MuteXt have been designed to retrieve mutations for specific superfamilies while MuGeX is targeted at proteins related to a specific disease. Despite these differences, these tools all solve several parts of the mutation extraction process similarly. Almost all tools, for example, use regular expressions to extraction mutations from text. We have studied the available tools to determine which aspects could be incorporated into Mutator.

Tool	*Database*	*Options*	*Year*	*Reference*
MuteXt	MEDLINE	KRGVF	2004	[17]
MEMA	MEDLINE	RGV	2004	[18]
Swiss-Prot	PubMed	KRGV	2007	[19]
CoagMDB	PubMed	KRGV	2008	[20]
Mutation Grab	PubMed	GVF	2007	[21]
MuGeX	MEDLINE	RG	2007	[22]
Vtag	NA	-	2004	[23]
Mtag	NA	KR	2006	[24]
OSIRISv1.2	MEDLINE	RGV	2006	[25,26]
MutationFinder	PubMed	-	2007	[27]
EnzyMiner	PubMed	KR	2009	[28]
Mutator	PubMed	KRGVF	2010	[29]

Table 2.2: List of mutation analyses tools. First two columns list the name of the tool and the database from which articles or abstracts for searching and scanning are retrieved. The following 5 columns list the capabilities of the tool: generation of lists for keyword searching, automatic retrieval of articles, grounding of mutations to proteins, validation of the grounding algorithm, and applicable to superfamilies. The publication date of the article accompanying the initial release of the tool and the reference are shown in the final 2 columns. Options: (K) Keyword searches, (R) Retrieval of articles, (G) Grounding of mutants to proteins, (V) Validation of mutants, (F) Support for protein or gene families.

1) MuteXt is part of the Molecular Class Specific Information System (MCSIS) platform, and offers an automated pipeline for retrieval of mutations for a complete protein superfamily. MuteXt was developed in close cooperation with the authors of the NuclearDB and GPCRDB MCSISes. The MCSIS database, used as backend, contains proteins and keywords of the target superfamily. The MEDLINE database is queried for relevant articles using query terms obtained from the MCSIS. Queries are generated by combining the protein family names with 'mutagenesis', 'mutant', and 'mutation'. MuteXt uses PubMed facilities to obtain full-text articles when possible. Mutations, gene names, and organism data are extracted from the articles using regular expressions. The first step in the grounding process is the selection of candidate sequences from the MCSIS database that match the sequence motif of the extracted mutation(s). Subsequently the corresponding sequence is determined by using the gene names and organism data extracted from the article. When multiple candidate sequences remain, the distance in words in the original text between the extracted terms and the mutation is used to select a sequence. The original design of MuteXt facilitates mutation retrieval for complete superfamilies. However MuteXt was designed to work with an MCSIS database as backend, and contains a lot of superfamily specific code. Therefore, MuteXt can only be used for the Nuclear Receptor and GPCR

superfamilies. MuteXt is no longer actively developed, but serves as a major inspiration for Mutator, similarly to the MCSISes serving as main inspiration for the 3DM platform at large. MuteXt was the first tool to utilize a superfamily system as backend for data storage. Many ideas from MuteXt are incorporated into Mutator such as the keyword generation, full-text article retrieval, usage of gene names, and organism data for grounding, and pre-selection of candidate sequences.

2) MEMA has been designed to extract protein and DNA mutations from all MEDLINE abstracts. MEDLINE contains only abstracts of articles classified as 'medical', which largely limits the application of MEMA to disease related proteins. Additionally, only abstracts which contain a HUGO [30] protein name were retrieved and processed. For Mutator we do not use HUGO but rely on protein names available through the UniProt and GenBank databases. Records from these databases contain lists of synonyms which generally include the main HUGO identifier in addition to other often used identifiers. Using the lists therefore increase the chances of recognizing protein identifiers in articles. MEMA assigns mutations retrieved from abstracts that contain only one gene directly to that gene. When multiple protein identifiers are observed, the corresponding protein is determined by evaluation of the text surrounding the mutation term. The MEMA authors validated and tested their method on several sets of articles. Recall and precision of the extraction of mutations and mutation-gene pairs from text was evaluated using a test set of 100 articles. Protein identifiers and mutations were manually extracted from these articles and compared with the results of MEMA. Extraction of mutations gave a recall of ~74% with a precision of 99% and for mutation-gene pairs a recall of only 35% with a precision of 93%.

3) Yip et. al describe a method to automatically update the references for variants in the Swiss-Prot database. All variants in the Swiss-Prot database have been manually extracted from the literature when available the effect of the mutant is also extracted. To keep the reference lists for each variant as up-to-date as possible a method was developed to validate existing, and retrieve new references for novel and existing variants from the literature. Protein names, gene names, and synonyms from the Swiss-Prot record are used to query PubMed for relevant articles. Any article that is already referenced by an existing variant is discarded. The abstracts and titles of the remaining articles are scanned using three regular expressions to extract mutations (table 2.3). A rudimentary sentence parser is also used to extract mutations: "For pattern 4, a sentence was considered as a match if it contained all of the following information: (1) a number (position of the variant); (2) a term describing the variation (e.g. variant, mutation, or polymorphism); (3) an amino acid in full letters, 3-letter code, or codon notation; and (4) a term descriptive of the position in the amino acid chain (e.g. position, located, situated). A sentence containing no term on position but mentioning two different amino acids was also considered as a match." Full-text articles

were used for preliminary tests to check if the workflow could be extended to include full-text articles. Although the tests indicated that full-text articles could be used the method proved to slow for inclusion in the regular workflow. New references for existing variants have been added automatically to the Swiss-Prot database, but novel variants could not be added automatically as the quality of detected mutations did not live up to Swiss-Prot's high accuracy standard. Articles containing novel variants were however tagged for later manual inspection. Unlike Mutator, Yip's method focuses on a high precision rate whereas Mutator values recall over precision. The regular expressions and search criteria used by Yip's et al. were studied and partially included in Mutator.

1	Single character abbreviations with position number	L45H, L-45-H, L(45)H
2	Three character abbreviations with position number	Leu45->His, Leu(45)His, CTT(45)CAT
3	Position number with three character abbreviations	45Leu->His, 45CTT>CAT, 45Leu→His

Table 2.3: Regular expressions used by Yip et. al to extract mutations from articles.

4) CoagMDB is a mutation database designed specifically for five sequentially highly similar serine proteases involved in blood coagulation. The keywords used to search PubMed for articles are based on Entrez gene names, alternative protein names and diseases from OMIM records associated to the genes. Articles that could not be retrieved automatically where retrieved manually. Mutations are not entered into the database automatically. Candidate sequences for mutation grounding are selected based on gene and protein names found in the text. The software uses regular expressions to locate sentences containing amino acids and residue positions. These sentences are extracted and parsed for mutation entities consisting of a wild type, mutant, and position. The wild type residues are matched with the candidate sequences to determine the most likely protein. CoagMDB uses three different numbering schemes for sequence selection: regular sequence numbering with or without the signal peptide, a chymotrypsin specific numbering scheme, and a sliding window approach. Regular sequence numbering checks if the wild type amino acid is present at the specified position in the sequence. As signal peptides can be included or excluded from the numbering at the discretion of the authors of the paper, the position is compensated for both possibilities. The chymotrypsin numbering of the proteases is also used to check if the wild type residue matches any of the candidate sequences. Alternatively, a sliding window is used to determine whether all mutations have a common offset in one of the candidate proteases. A web interface is used to show the results from the parser to an administrator who can either accept or discard the mutation(s). Using alternative superfamily specific numbering schemes will be added to Mutator in the future. To prevent having to include superfamily specific code into Mutator the numbering schemes will be stored in the 3DM database and be editable by the user. A sliding window to find mutations in sequences will

not be incorporated into Mutator in the foreseeable future as using a sliding window with the vast amount of sequences available in the backend database would significantly increase the number of false-positives.

5) Mutation Grab was tested on three protein families: G protein-coupled receptors, tyrosine kinases, and ion channel transporters. These three families were selected as they are the three major targets for drug development, Mutation Grab has been designed to be applicable to all superfamilies. Articles were selected differently for the superfamilies. The tGRAP [31] database was used to identify articles for the GPCR family. For the other two families the family name in combination with "protein mutation" was used as PubMed query. In all three cases the articles were curated manually and 100 articles were selected to be used as training set, and 100 articles were selected as test set. The remaining articles were not used for the analyses. Mutations, protein names, and organism names were extracted using regular expressions. Regular expressions to extract protein and organism names where based on dictionaries assembled from Swiss-Prot and Entrez gene proteins related to the protein family. Candidate sequences were retrieved from Swiss-Prot using the organism and protein name. Further selections have been made by fitting the wild type residues from the mutation on the sequence. Isoforms and signal peptides were taken into account. When only one candidate protein remained the mutation was assigned to that protein. When multiple candidates remained a graph bigram metric was used. Graph bigram is a distance metric to determine the likelihood two words are connected. Graph bigrams include frequency and positional data for all terms extracted from a text. For example, if "human GLA" is found at the beginning of the article and GLA itself is found close to the location of the mutant in the text, the graph bigram metric will associate human with the mutation in addition to GLA. Both terms can then be used to select a protein.

6) MuGeX retrieves all abstracts from MEDLINE that contain either the word "mutation" or "polymorphism" and extracts amino acid mutations from these abstracts. A preprocessing step is used to extract authors, title, and journal from the abstract and store them in an easily retrievable format. The title and content from the abstract are first split into sentences and subsequently into words for later processing. A set of 20 regular expressions all based on the [amino acid][position][amino acid] format are used to extract mutations. Machine learning techniques are used to filter out nucleotide mutations, strain names, and cell lines from the results that matched any of the regular expressions. For the grounding step a dictionary is created using HUGO gene names and aliases. Because a standard nomenclature for gene names is lacking the terms in the dictionary are preprocessed in order to normalize the gene names. For gene symbols all hyphens are replaced by white space, and all parenthesized material is removed. For gene names, all parenthesized materials are also removed, all punctuation characters are replaced by white space, and all uppercase characters are lower

cased. Grounding of mutations to genes follows 4 simple rules based on the method used by MEMA: The gene selection rules are: 1) the abstract contains only one gene; 2) the sentence the mutation was found in contains only one gene; 3) the closest gene based on proximity in the text when multiple genes are found in the sentence containing the mutation; 4) the first sentence of the abstract; Testing of the MuGeX method on a selected set of 231 MEDLINE abstracts containing 472 mutations revealed recall and precision rates of over 85%. The filtering of false-positives using the machine learning based disambiguation module resulted in a higher precision at a slight cost to the recall. An overall test of MuGeX performance was done using Alzheimer's disease. Alzheimer's was selected as several databases with mutation are available allowing for validation of the results. The tested showed that MuGeX misses most mutations due to the gene names not being available from HUGO. A website was made available to search for mutations in predefined enzyme classes using MuGeX. Using a test set of 231 articles containing 334 distinct gene names the MuGeX algorithm had a recall rate of only 56%. Most missed gene names are simply lacking from the HUGO database, the remaining cases are mainly spelling variants that could not be resolved by the normalization algorithm. The authors suggest the use of extended dictionaries with all possible versions of the HUGO gene names to increase recall. Mutator does not solely use the HUGO gene names for precisely this reason. A website is available were the MuGeX results can be queries for mutations by gene or disease.

7) VTag and MTag are named entity taggers. VTag targets mutation and deletions, called "variation events" in the article hence VTag, whereas MTag was designed to tag malignant cancer references such as neuroblastoma. Named entity tagging is a process where predefined entities are located and highlighted (tagged) in a text. Both tools use the machine-learning method Conditional Random Fields (CRF) that recognizes text entities. CRF is used to assign types and weights to tokens retrieved from abstracts. For example, VTag annotates several tokens from the sentence, "All cases with K-ras codon 12 mutations were found to be G to T transversion": type: transversion, location: codon 12, wildtype: G, mutation: T. Rather than retrieving articles and extracting mutations, both tools are restricted to tagging occurrences for manual reviewers. Neither VTag nor MTag contains article retrieval or mutation grounding functionality which is left to curators or other tools.

8) OSIRIS uses a Named Entity Recognition (NER) module to extract variants from bio-medical literature. These variants are mapped on dbSNP entries with the use of a custom mutation database, HgenetInfoDB. Every MEDLINE abstract is retrieved from a local MEDLINE mirror and annotated using the ProMiner [32] NER by default. OSIRIS has however been designed to work with any NER capable of grounding proteins to NCBI Gene identifiers. Results from both steps are stored in a local database and available from a website. Search options are available using both genes and MESH terms.

9) MutationFinder. Caporaso *et. al* aimed at providing a freely available implementation of a mutation extraction tool and a dataset of annotated abstracts which could be used as reference for validation. A partial reimplementation of MuteXt was used as basis for MutationFinder. The authors of MutationFinder decided not to include the sequence verification step in the grounding algorithm to increase recall at the expense of precision. Six changes were made to the mutation extraction module as compared to MuteXt. 1) w*N*m format mentions with one-letter abbreviations must have $N > 9$; 2) w*N*m format mentions with one-letter abbreviations must appear in upper-case letters; 3) Wild-type and mutant residue/base identities must not be the same; 4) MutationFinder specifies patterns incorporating non-alphanumeric characters, whereas the baseline system removes non-alphanumerics; 5) MutationFinder identifies mutations described in natural language (as opposed to completely abbreviated formats) with specific patterns, whereas the baseline system uses a heuristic to match these mentions; 6) MutationFinder splits text on sentences and applies its regular expressions to each sentence, whereas the baseline system splits both on words and sentences and applies different regular expressions to each; Several of these updates to the MuteXt algorithm were also integrated into Mutator. Mutations on the first 9 positions of a protein receive a penalty score in the grounding algorithm of Mutator. Like MutationFinder non-alpha numerical characters are preserved. However, the conversion process of PDF documents to text in Mutator converts most non-alpha numerical characters to ASCII equivalents. Extracting mutations from natural language is also implemented in Mutator. The capitalization of wildtype and mutant as specified in adaptation 2, and the skipping of silent mutations in adaptation 3 are not included in Mutator as we focus on recall over precision. Abstracts of 813 articles were annotated with in total 1,515 mutations. 508 abstracts were used as test set, and 305 as reference set. The entire set of annotated abstracts was made publically available to be used as golden standard set for the validation of mutation extraction tools. MutationFinder is freely available in three different programming languages: python, perl, and java.

10) EnzyMiner was developed to extract and classify protein mutations from PubMed abstracts. A PubMed query using the enzyme name and "mutation" is used to find relevant articles. The Mutation Extraction module from MuGeX [22] is used to retrieve mutations from these abstracts. As EnzyMiner only handles protein mutations, the Disambiguation module from MuGeX is used to discard abstracts containing only genetic mutations, or terms wrongly identified as mutations such as strain names. Three separate classifiers are constructed using RainBow [33] to separate the abstracts into classes. The first classifier distinguishes between diseases and non-diseases related abstracts. The non-diseases related abstracts are separated into change and no change classes by the second classifier. The final classifier is used for the change abstracts which are divided into stability and catalytic classes. All classifiers use a Probabilistic Indexing algorithm although different stemming and tokenizers were selected for each classifier. A set of 194 abstracts containing mutation for the amylase or

lipase enzyme classes was used to train the classifiers. 155 abstracts were used as training set with the remaining 39 abstracts used as test set. To test if training the classifiers on a single enzyme family results in a generic classifier the classifiers were retrained and test on different enzyme classes. The results indicated a high accuracy from which the authors concluded that abstracts from the target enzyme class do not have to be part of the intial training set for EnzyMiner to perform accurately. EnzyMiner has several drawbacks such as the inability to assign type or direction to mutations in the change classes. For example, EnzyMiner cannot determine whether a mutation has an effect on Ph or temperature resistance and if so whether the mutation increases or decreases the stability. Additionally, EnzyMiner does not contain a grounding algorithm. A distinction can therefore not be made between mutations from different proteins in the same abstract. Results from EnzyMiner are visualized using a website. A selection can be made between disease or non-disease mutations for a number of enzyme classes. For each mutation a sentence with highlighted keywords selected by the classifier is shown together with the PubMedID and mutation(s) and publication year.

11) Mutator is mainly inspired by MuteXt. The MCSIS platform is the predecessor of 3DM with MuteXt similarly being the predecessor of Mutator. Mutator uses a similar architecture and workflow compared to MuteXt. A superfamily database serves as backend for storage of sequences and mutations. Keyword lists are generated using protein and amino acid annotations from the database. Article retrieval is automated and regular expressions are used to extract mutations from text. The grounding algorithm used in Mutator is also based heavily on the MuteXt implementation. From the other tools we incorporated additional regular expressions to extract more mutations. OMIM was added as a source for PubMed search keywords to find relevant articles. A future addition to Mutator will be the usage of custom numbering schemes defined in the 3DM database for the selection of candidate sequences. The following chapter describes the various steps in the mutation retrieval process as implemented by Mutator in more detail.

2.6.2 Challenges

Mutation related data can be valuable in the study of protein function. Obtaining a relevant set of articles and mutations for a protein of interest is a time consuming process. Articles must be selected, retrieved, and parsed to extract mutations. Mutations must be linked to the correct protein, a process called grounding. Each step in the retrieval process is faced with challenges. In Mutator we recognize six distinct steps that will be described in the following sections.

1) Article selection. PubMed is the main publicly available online source for articles, currently containing over 20 million references. Only a small fraction of these articles are related to a protein or superfamily of interest and an even smaller fraction contains mutation data.

The list of articles to process should be as small as possible since retrieving, converting, and scanning articles is a time-consuming process. Articles are selected by querying the PubMed database using the NCBI Entrez Programming Utilities [10]. Protein names, gene names, functional names, diseases, and superfamily keywords are used to restrict the set to articles concerning the superfamily of interest. Mutation related search keywords, such as “mutant”, “mutagenesis”, and “mutational” are used to limit the number of articles that are linked to the field of interest but do not contain any mutational data.

2) Article retrieval. PubMed search results contain links to websites of publishers or journals from which the article can be retrieved. The links often do not direct the user to the article itself but rather to the website of a publisher or journal. Fully automatic retrieval of articles from these websites is a difficult process. Publisher websites often use javascript obfuscation to hide files, redirect the user several times, or try to open the article in the browser itself. Additionally the website layout and steps required for the download process are redesigned often. Several options are available to retrieve the full-text PDF file. The location of the PDF file itself is determined from the HTML code of the page. Depending on the publisher website either a web-browser, or downloader will be used to obtain the file. Manual download of articles will almost always succeed, however this is not a viable option as hundreds of articles might be relevant for the protein or superfamily of interest. Due to the constant redesign of web-sites article retrieval requires occasional manual supervision and intervention to remain able to obtain all required articles. A further limited for the electronic availability of articles are age and subscriptions. Older articles are often only available through libraries whereas newer articles may not be accessible due to missing journal subscriptions.

3) Article conversion. Before mutations can be extracted from articles a conversion step is required for all non-text formats. Most articles are available in PDF format [34] which was a proprietary Adobe format until 2008. Several tools are available for the conversion of PDF to text. Either open source tools such as pdftotext (xpdf http://www.foolabs.com/xpdf), poppler (http://poppler.freedesktop.org/), and utopia documents [3] or commercial tools such as simpo pdf converter (http://www.simpopdf.com/). Common conversion problems of these programs are the conversion of non-standard characters such as Greek letters used in gene names and the contents of tables and images. Any information present in images is currently unavailable as none of the tools is capable of converting images to text. Data in tables can be used but tables may be placed either horizontally or vertically and the column and row structure is often lost in conversion. Mutator uses the pdftotext utility from the poppler suite.

4) Mutation detection. The terminology used to describe a mutation varies between journals and research fields due to a lack of commonly accepted nomenclature. A mutation may be written in short form such as A23F, p.A23F, or Ala23Phe, or in a more descriptive format such as "the alanine to phenylalanine mutation on position 23" or "resulting in a phenylalanine on position 23". The sequence numbering used to indicate the position of the mutant in the protein also differs from article to article. Initiator methionine and signal peptide may or may not be included in the numbering scheme and different isoforms often have different lengths and numbering schemes. Additional obfuscation results from numbering schemes commonly used in some superfamilies. Antibody sequences, for example, are often numbered using Kabat [35], Chotia [36], or a derived numbering scheme [37]. These schemes define numbers for conserved residues in the antibody family. Insertions required for antibodies with longer sequences are specified with the addition of a character to the position after which the insertion occurs. L82cA is therefore a valid identifier for a leucene to alanine mutation of the third residue between positions 82 and 83. The possible numbering schemes and amino acid identifiers result in an almost infinite number of combinations. Using a combination of numbering schemes and identifiers most terms can therefore be interpreted as a valid mutation. An additional problem is the ambiguity of some of these terms. For example, a C23G mutation may either indicate an amino acid or a nucleotide mutation. Some journals enforce rules on mutation formats such as an obligatory prefix indicating whether the mutation is an amino acid (p.) or nucleotide (c.). However, these rules themselves have not been standardized and differ from journal to journal. Even apparently valid and straight forward terms such as 3A-G may indicate different entities e.g. a mutation at position three or the abbreviation of 3-(arylmethylidene)aminoxy-(3a-g, 4a-g) acid. Additional false positives might even occur from the name of the superfamily. Mutator results for the P450 superfamily contained an abnormally high number of proline mutants on position 450. These mutations were almost invariably variations on the name of the superfamily which had been mistaken for mutations and ground successfully to sequences which actually do contain a proline on position 450.

To retrieve mutations from full-text articles two methods have been commonly used: regular expressions and Natural Language Parsing (NLP). Regular expressions are structured search terms that can be used to find and retrieve sections from text that match the expression. For example, the expression [A-Z][0-9]+[A-Z] will search for a capitalized character followed by one or more digits, indicated by the plus, followed by another capitalized character. Regular expressions can be structured to match a wide variety of mutation formats. However, the expressions are rigidly evaluated. For example, the given expression will not match mutations when a space is inserted between the amino acids and the position, when the amino acids are written using their three letter abbreviations or when lower case characters have been used. All of these problems can be solved by adapting the expressions to allow spaces, mixed characters or various other exceptions. This however leads to an ever growing set of expressions to allow for increasingly obscure corner cases. The

number of retrieved mutations will increase at the expense of an increased number of false-positives. An alternative method for mutation retrieval is Natural Language Parsing. NLP based methods 'read' the text and attempt to extract the meaning by analysing sentences. NLP has been a promising technique for at least 20 years but actual applications in mutation retrieval are scarce [38]. The main problem hampering the wider adoption of NLP is the incredible diversity of human languages. A sentence describing a mutation in a gene can be written in hundreds of different ways. Even splitting text into sentences is difficult as not every dot indicates a new sentence, and not every sentence is necessarily finished with a dot. A technique which may solve both the rigidity problem of regular expressions and the interpretative problem of NLP is RDF. RDF parses text and creates triples which assign meaning to words. For example GLA [is a] gene, where GLA is the term retrieved from the article which is linked to the gene keyword with which it has an "is a" relationship. RDF is somewhere midway between NLP and regular expressions in complexity. An increasing number of public databases are offering RDF annotated versions of their data. In the future, RDF might therefore be used to obtain data or annotations of data for usage in text mining.

Mutator uses a series of regular expressions to extract the mutations from text. These expressions are partially taken from publications on other mutation extraction tools such as Yip *et al.* and partially developed and adapted from missed mutations in manually verified articles.

5) Mutation grounding. The protein in which a mutation was found is not always mentioned explicitly in the article. Instead of a protein identifier such as a UniProt accession code, only a gene name or function in combination with a species is often used to indicate the protein. Changes and updates in the protein databases affect identifiers used in (especially older) articles when the submitted sequences are updated, or removed entirely. Proteins in UniProt that are obtained by translating nucleotide sequences in, for example, GenBank are updated or removed when the original genomic sequence is found incorrect or incomplete. Formerly popular databases may also have disappeared due to lack of funding, or public interest [39].

Mutator uses dictionary based searches to retrieve organism- and gene name data from articles to assist mutation grounding. An initial selection of sequences is made based on sequence motif searches using the wild type residues from the mutations obtained from the articles. For example, if mutation A23D and F24E were retrieved from the article text, all sentences with motif 23A,24F are retrieved from the 3DM database. Dictionaries are created based on the organism name and on the gene name data available for the selected candidate sequences. Both the scientific and the common name are used for the organism dictionary. To ground the mutation, all candidate sequences are evaluated individually and graded using a scoring scheme. For each of the mutants found in the text that matches with the candidate sequence, one point is added. An additional point is added for matching organism or gene names, with 0.5 points added when the organism or gene name is found

in the title. When only the genus of the organism is found 0.5 is added to the score instead of 1.0. A penalty score is subtracted when more then 15 mutants were found in the text to prevent that articles with many mutations match a sequence by chance alone. This penalty is the number of mutations divided by 15. Articles also often contain molecular formulas of various compounds or abbreviations of genes or species which can easily be mistaken for mutations because numbers in these strings are often below 10. Mutations in the first 10 positions of a sequence therefore, receive a penalty of one point. When multiple candidate sequences are available for a mutation the score is calculated individually for each candidate. A minimum score of 3.5 is required for a candidate sequence to prevent false-positives matches. When multiple candidate sequences are available the mutation is grounded to the highest scoring candidate. When the score is a tie, Mutator cannot determine which of the candidate sequences is the actual sequence used in the original study. Rather than selecting one of the sequences arbitrarily, the mutation is grounded to all candidate sequences. The task of curator is left to the user of the system who will have to disambiguate the grounding. This is in line with the 3DM wide strategy to always include data even when this might impact accuracy.

6) Superfamilies. Manual mutation retrieval for a single protein of interest could in principle be performed in an acceptable time-frame. However, single protein PubMed queries will miss many articles that do not describe the protein itself but instead deal with closely-related superfamily members. These articles are also relevant for the single-protein case as data related to these mutations can often be transferred to the protein of interest. This greatly increases the list of articles that potentially contain useful mutational information. The size of protein superfamilies ranges from a few hundred to tens of thousands of distinct proteins, which means thousands of articles may match the query terms. Manual retrieval of mutation from relevant articles is therefore infeasible due to the time required to retrieve, scan, ground, and verify mutations from so many articles.

2.7 DNA Diagnostics

Many diseases of many different types are known. Diseases for which a genetic origin has been established are called genetic disorders. The human genome contains at least 10,000 protein coding genes leading to many potential candidates for deleterious mutations. The total number of genetic disorders is affected both by genes with a duplicate function since defects in one gene can be compensated for by another gene, and composite genes where defects in different regions of a single gene give rise to separate genetic disorders. OMIM lists over 3,000 disorders with a known molecular basis as of spring 2012. Diagnosing patients suffering from genetic diseases however remains difficult for several reasons: the overlap in phenotypes between different diseases, the large amount of variance found in the human genome, and the lack of a one-to-one mapping between diseases and genes. The research

field of DNA diagnostics focusses on diagnosing (and treating) patients suffering from genetic diseases. For diagnosis the underlying genetic defect must be located. Treatment of genetic disorders requires determining the effects of the disorder on the protein. The following sections will discuss datatypes, databases and tools that can potentially be used for diagnosing genetic disorders.

2.7.1 Diagnostic Data Types

Many different data types can be used to determine the pathogenicity of unknown variants (UVs) such as solvent accessibility, bumps, conservation, and active site residues. The scripts used to obtain this data are described in section 2.4.11 unless indicated otherwise. Some of the available data types for variant analysis are:

1) Genetic variants. Many different projects study genetic variants such as mutagenesis experiments, GWAS studies, patient trials, etc. Variants can be used for two purposes: 1) when an effect is available, a direct link can be made between the variant and a phenotype; 2) variants can be used to determine the conservation of genomic locations. Discovered variants are often published using articles and sometimes deposited in variant databases (see chapter 2.7.2 for an overview). We designed Mutator (chapter 1.4 & 2.6) to automatically retrieve mutations from articles. Variants stored in databases are retrieved using custom scripts.

2) Contacts. Ligand contacts, dimer interfaces, salt-bridges, and charge clamps are vital for the stability, folding, and functional properties of proteins. Disrupting these contacts often results in unstable or inactive proteins. Variants at such positions are therefore more likely to be pathogenic than variants at positions that are not involved in functionally important contacts. Ligand contacts and dimer interfaces are retrieved using WHAT IF; charge clamps are determined using a custom script.

3) Amino acid variability and conservation. Amino acid variability and hence conservation can be determined using alignments. Alignments are constructed of naturally occurring proteins and therefore reflect the variability allowed in nature. The occurrence of an amino acid at a position can therefore be used as an indication of the viability of the variant. Alignments often reflect millions of years of trial and error and therefore variants that do not occur in natural proteins are most likely detrimental. These variants are thus more likely pathogenic compared to variants that do occur in nature. Alignments can also be used to determine the conservation of a position. Highly conserved positions are often involved in important structural of functional properties of the protein. Variants at conserved positions are therefore more likely to be pathogenic then variants at non-conserved positions.

4) Solvent accessibility. Solvent accessibility is the degree to which a residue is exposed to solvents. Solvent accessibility can be determined using protein structures. Inaccessible residues are often buried within the protein, and potentially involved in intramolecular contacts such as hydrogen bonds. Variants at buried positions rarely fit in the available space due to differences in size, charge, and preferred side-chain orientation. Additionally, smaller variants that do fit might disrupt intra-molecular contacts required for folding. For Fabry's disease, for example, Garman [40] has shown a strong correlation between pathogenic mutations and buried residues in the corresponding GLA gene.

5) The number of bumps caused by a variant. A bump is defined as atoms from two residues taking up the same space in the protein. This usually results from buried mutations where a smaller residue is replaced with a larger residue which does not fit in the available space. Bumps in a protein will either force a different folding, or disrupt folding entirely. The number and severity of bumps caused by a variant is therefore related to the likelihood the variants effect will be pathogenic.

6) Amino acid annotations. Amino acids can be involved in many functions in the mature protein such as active sites, post-translation modifications, trans-membrane regions etc. Variants at these positions often disrupt the function or regulation of the protein. Annotations of interest are phosphorylation sites, glycosylation sites, active sites, domains, etc.

7) Structural location. The location of a variant can also be used to determine pathogenicity. An amino acid can be located in a helix or beta-sheet or be part of a trans-membrane domain. Different amino acids are preferred in these regions and some variants would result in disruption of the secondary structure.

8) Amino Acid substitution matrices. Manyn different amino acid substitution matrices are available, such as BOLSUM and Dayhoff (see chapter 1.2). The matrices are constructed by observing the frequency of mutations from each amino acid to all other amino acids. Amino acid mutations which occur more frequently are given a lower penalty compared to mutations which occur more rarely. Example matrices are Grantham [41] and BLOSUM [42]. Scores from these matrices are used to indicate the magnitude of the amino acid substitution and thereby the chance that the mutant causes a genetic disease. These matrices however, only reflect the magnitude of a specific substitution without taking the environment of the amino acid into account. Therefore, these matrices are much less predictive of pathogenicity compared to the sequence and protein environment based criteria, and should only be used as a last resort if other data yield ambiguous results.

2.7.2 Variant Databases

Prior data on UVs can be used to assist in determining variant pathogenicity. Several databases are available containing data on known protein and DNA variants. Examples of such databases that contain generic human variants are OMIM [43], dbSNP [44], HGVBase [45], HGMD [46], and Swiss-Prot [47]. Specialized variant databases are available for some proteins or diseases, such as the tumor linked p53 gene [11].

1) OMIM. The Online Mendelian Inheritance in Man (OMIM) database contains data on diseases in humans. OMIM lists all known Mendelian diseases and contains data on over 12.000 genes. For each disease the relevant articles are listed and, if available, the genetic background is described. Population and demographic data on disease severity and impact are also included. Variant data is manually retrieved from literature, an automatic article scanner selects publications for review. OMIM can be accessed through the NCBI search website and is available both in academic and commercial versions.

2) dbSNP. dbSNP is a database that contains both natural and pathogenic variants for several species including human, chicken, and mouse. The population where the variant was found and occurrence in the population are available for some SNPs. Variants linked to disease related genes are linked with the corresponding OMIM entries. Variant data can be entered into dbSNP via external submission methods.

3) HGVBase. The Human Genome Variation database (HGVBase) is a central repository for mutation collection efforts by members of the Human Genome Variation Society (HGVS). The database aims to provide a non-redundant overview of all known and suspected variants in the human genome. High quality of the database contents is guaranteed by extensive manual curation of submitted sequences. Publically available SNP databases such as dbSNP are scanned and variants passing quality control are incorporated into the system. In 2011 the database was renamed to GWASCentral to emphasize the shift in focus towards collection of variants from genome wide association studies (GWAS).

4) HGMD. The Human Gene Mutation Database contains a collection of pathogenic variants in human. Variants are retrieved from literature manually. The manual extraction allows for additional data on variants such as phenotype to be incorporated into the database. Variants in transcription regions are also included. Publications are linked for each variant, however each unique variant is only reference once. The HGMD is run by the University of Cardiff and commercially exploited by Biobase. Access to the complete database is only available for users with a paid subscription.

5) Swiss-Prot. Swiss-Prot records list both natural and pathogenic variants. Separate variant pages are available which list the variants with their effect: disease, polymorphism or unclassified. 3D models are generated whenever possible.

2.7.3 Variant Calling Tools

Variant databases are primarily useful for the study of known variants and cannot be used to predict the effects of UVs. Several tools have been developed to assist in variant analysis and pathogenicity prediction. Examples of these tools are ProCMD [48], HOPE [49], mSTRAP [50], SIFT [51], PolyPhen [52] and Alamut [53].

1) ProCMD. ProCMD, the Protein C Mutation Database, focuses on mutations in protein C. Protein C is an anticoagulant plasma serine protease that plays an important role in controlling inflammation and cell proliferations. Mutations in protein C have been linked to venous thrombosis. The database contained a total of 195 mutations upon publication: 21 variants were retrieved directly from in-house patient data, the remaining mutations were obtained from the HGMD, Swiss-Prot Variants, and an existing database of protein C gene mutations. Alignments with paralogues and homologues sequences are used to determine amino acid conservation. A structural model can be generated for most mutants for additional study. ProCMD does not address the fitness or pathogenicity of mutated proteins, is only applicable to protein C, and the database does not appear to have been updated with new mutations since its inaugural publication in 2007.

2) The mutation extraction and STRucture Annotation Pipeline, mSTRAP, is an automatic workflow to collect mutations from articles and visualize the mutations in a structure. Articles are retrieved from PubMed using user provided keywords as search terms. mSTRAP automatically downloads and converts the articles to plain text before extracting mutations and protein names. A single regular expression is used to extract mutation formatted as A23G while allowing for the use of three-letter amino acid abbreviations, and a dash between the wildtype amino acid and position. A list of all protein names available from Swiss-Prot is used to extract protein identifiers from the text. Mutations and proteins are matched based on occurrence in the same sentence. BLAST searches are used to determine the closest structural homolog for each protein. At most three homology models are built and evaluated using a scoring algorithm. jMol is used to present the user with the selected structure. All mutations mapped to the protein are visualized in the structure and the sentence from which the mutant was extracted is shown.

3) SIFT can be used to Sort Tolerant From Intolerant mutations in proteins or genes. For a given input sequence SIFT uses PSI-BLAST to construct an alignment of homologous

sequences. Additional precautions are made to prevent the alignment consisting of very close homologs which might occur for example for viral strains which are often sequenced with only one or two mismatches multiple times. A prediction on whether the mutation is damaging is made solely on the conservation of the position in the alignment. Highly conserved positions are deemed evolutionary important and therefore unlikely to allow substitutions. Validation of SIFT results on a test set of patient data resulted in ~70% of mutants correctly predicted as damaging.

4) PolyPhen predicts mutant pathogenicity using a combination of eight sequential and three structural parameters. The parameters include the number of residues in the generated alignment, the change in residue side chain volume, and the change in accessible surface of buried residues. An alignment is automatically generated using sequences selected by a clustering algorithm. For each mutation a Naïve Bayes posterior probability that the mutation is damaging is calculated. An estimate of the false positive and true positive rates for the variant are also determined. PolyPhen2 was evaluated using two training sets, both consisting of annotated benign and pathogenic mutations. Pathogenic mutations were retrieved from OMIM and UniProt and benign sequence variants were retrieved from UniProt. Testing PolyPhen2 on these training sets showed a good to average true positive rate of 92% and 73% for the test sets but a false positive rate of 20%.

5) Alamut is a knowledgebase system similar to 3DM but mainly focussed on genes instead of proteins. Variants are retrieved from external databases such as HGVBase and presented using a standardized nomenclature. Alamut integrates with several external databases and tools such as PubMed, Ensembl, and Uniprot. Mutations can be evaluated using external prediction tools such as SIFT and Ployphen. A search engine is available to query PubMed abstracts for mutations, however no complete text-mining tool is incorporated in Alamut.

6) HOPE, our own pathogenicity predictor is aimed at biomedical researchers. A website is available where a sequence and mutant can be entered for analysis. Data for the protein of interest is gathered from UniProt and PDB, and calculated using WHAT IF and DAS using webservices. The advantage of webservices is the distributed nature which results in relatively few in-house tools that need to be maintained. Together with the analysis, HOPE also provides snapshots and movies of the variant in the structure or a homology model when possible. Project Hope is described in more detail in chapter 4.

2.8 Wrap-up

The various subdisciplines of the life-sciences are producing an impressive amount of data on a daily basis. A lot of this data is published and made publicly available for reuse in

other projects. Large public database offer access to structures, genes, proteins, annotations, literature, and many other data types. Many specialized data-sets are also available such as a large number of transcriptomics experiments on genes with a multitude of different variables. New research projects often start by collecting the literature available for the protein/gene/function/disease of interest. A thorough literature study however might result in hundreds of articles which might be relevant for the new project. A similar problem arises in the next phase of the project: obtaining relevant existing data. Obtaining both literature and data is complicated by the ambiguity of data relevance. Data which is not directly related to the field of research or the protein of interest may be invaluable in solving the original hypothesis.

Bioinformatics and systems biology are fields that aim at incorporating all available data into databases which can be queries easily by biologists. A choice must however always be made between the capabilities of the hard- and software, breadth of data to store, maintainability, and ease of use. Specialized systems, such as 3DM, focus on a subset of all data and try to solve a limited number of problems rather than trying to be the swiss-army knife of the life sciences. Such an army knife would have so many functions and utensils it would probably be unpractical in actual day-to-day use. 3DM systems for example are not aimed at solving genomics or transcriptomics problems, and (currently) do not even store most of the data that would be necessary to do this. The previous two chapters have described the scientific and technological background, some applications, and evolution of the 3DM platform. Several examples have been presented to show the generic capabilities or highlight distinctive features of the platform. The following chapters give a more detailed overview of 3DM itself and the techniques used, the kind of problems 3DM can be used to solve, and the applications it has been put to.

References

1. G. Vriend, WHAT IF: a molecular modeling and drug design program, J Mol Graph, 8 (1990) 52–56, 29.
2. E. Krieger, YASARA, n.d.
3. T.K. Attwood, D.B. Kell, P. McDermott, J. Marsh, S.R. Pettifer, D. Thorne, Utopia documents: linking scholarly literature with research data, Bioinformatics, 26 (2010) i568–574.
4. A.L. Cuff, I. Sillitoe, T. Lewis, A.B. Clegg, R. Rentzsch, N. Furnham, M. Pellegrini-Calace, D. Jones, J. Thornton, C.A. Orengo, Extending CATH: increasing coverage of the protein structure universe and linking structure with function, Nucleic Acids Res., 39 (2011) D420–426.
5. R.P. Joosten, T.A.H. te Beek, E. Krieger, M.L. Hekkelman, R.W.W. Hooft, R. Schneider, C. Sander, G. Vriend, A series of PDB related databases for everyday needs, Nucleic Acids Res., 39 (2011) D411–419.
6. G. Vriend, C. Sander, Detection of common three-dimensional substructures in proteins, Proteins, 11 (1991) 52–58.
7. R.C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Research, 32 (2004) 1792–1797.
8. The PHP Group, PHP, (n.d.).
9. World Wide Web Consortium (W3C), XML Path Language (XPath) 20, (n.d.).
10. E.W. Sayers, T. Barrett, D.A. Benson, E. Bolton, S.H. Bryant, K. Canese, V. Chetvernin, D.M. Church, M. Dicuccio, S. Federhen, M. Feolo, L.Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D.J. Lipman, Z. Lu, T.L. Madden, T. Madej, D.R. Maglott, et al., Database resources of the National Center for Biotechnology Information, Nucleic Acids Res., 38 (2010) D5–16.
11. M. Olivier, R. Eeles, M. Hollstein, M.A. Khan, C.C. Harris, P. Hainaut, The IARC TP53 database: new online mutation analysis and recommendations to users, Hum. Mutat, 19 (2002) 607–614.
12. S.R. Engel, R. Balakrishnan, G. Binkley, K.R. Christie, M.C. Costanzo, S.S. Dwight, D.G. Fisk, J.E. Hirschman, B.C. Hitz, E.L. Hong, C.J. Krieger, M.S. Livstone, S.R. Miyasato, R. Nash, R. Oughtred, J. Park, M.S. Skrzypek, S. Weng, E.D. Wong, K. Dolinski, et al., Saccharomyces Genome Database provides mutant phenotype data, Nucleic Acids Res., 38 (2010) D433–436.
13. Network Working Group, RFC 4180: Common Format and MIME Type for Comma-Separated Values (CSV) Files, (n.d.).
14. S. Velankar, G.J. Kleywegt, The Protein Data Bank in Europe (PDBe): bringing structure to biology, Acta Crystallographica Section D Biological Crystallography, 67 (2011) 324–330.
15. R. Hooft, G. Vriend, WHAT IF Manual, (n.d.).
16. R.G. Cote, P. Jones, L. Martens, S. Kerrien, F. Reisinger, Q. Lin, R. Leinonen, R. Apweiler, H. Hermjakob, The Protein Identifier Cross-Reference (PICR) service: reconciling protein identifiers across multiple source databases, BMC Bioinformatics, 8 (2007) 401.
17. F. Horn, A.L. Lau, F.E. Cohen, Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors, Bioinformatics, 20 (2004) 557–568.
18. D. Rebholz-Schuhmann, S. Marcel, S. Albert, R. Tolle, G. Casari, H. Kirsch, Automatic extraction of mutations from Medline and cross-validation with OMIM, Nucleic Acids Res, 32 (2004) 135–142.

19. Y.L. Yip, N. Lachenal, V. Pillet, A.-L. Veuthey, Retrieving mutation-specific information for human proteins in UniProt/Swiss-Prot Knowledgebase, J Bioinform Comput Biol, 5 (2007) 1215–1231.
20. R.E. Saunders, S.J. Perkins, CoagMDB: a database analysis of missense mutations within four conserved domains in five vitamin K-dependent coagulation serine proteases using a text-mining tool, Hum. Mutat, 29 (2008) 333–344.
21. L.C. Lee, F. Horn, F.E. Cohen, Automatic extraction of protein point mutations using a graph bigram association, PLoS Comput. Biol, 3 (2007) e16.
22. M. Erdogmus, O.U. Sezerman, Application of automatic mutation-gene pair extraction to diseases, J Bioinform Comput Biol, 5 (2007) 1261–1275.
23. R.T. McDonald, R.S. Winters, M. Mandel, Y. Jin, P.S. White, F. Pereira, An entity tagger for recognizing acquired genomic variations in cancer literature, Bioinformatics, 20 (2004) 3249–3251.
24. Y. Jin, R.T. McDonald, K. Lerman, M.A. Mandel, S. Carroll, M.Y. Liberman, F.C. Pereira, R.S. Winters, P.S. White, Automated recognition of malignancy mentions in biomedical literature, BMC Bioinformatics, 7 (2006) 492.
25. J. Bonis, L.I. Furlong, F. Sanz, OSIRIS: a tool for retrieving literature about sequence variants, Bioinformatics, 22 (2006) 2567–2569.
26. L.I. Furlong, H. Dach, M. Hofmann-Apitius, F. Sanz, OSIRISv12: a named entity recognition system for sequence variants of genes in biomedical literature, BMC Bioinformatics, 9 (2008) 84.
27. J.G. Caporaso, W.A. Baumgartner Jr, D.A. Randolph, K.B. Cohen, L. Hunter, MutationFinder: a high-performance system for extracting point mutation mentions from text, Bioinformatics, 23 (2007) 1862–1865.
28. S. Yeniterzi, U. Sezerman, EnzyMiner: automatic identification of protein level mutations and their impact on target enzymes from PubMed abstracts, BMC Bioinformatics, 10 Suppl 8 (2009) S2.
29. R.K. Kuipers, H.-J. Joosten, R.H. Lekanne dit Deprez, M.M. Mannens, P.J. Schaap, Novel tools for extraction and validation of disease-related mutations applied to Fabry disease, Hum. Mutat, 31 (2010) 1026–1032.
30. T.A. Eyre, F. Ducluzeau, T.P. Sneddon, S. Povey, E.A. Bruford, M.J. Lush, The HUGO Gene Nomenclature Database, 2006 updates, Nucleic Acids Research, 34 (2006) D319–D321.
31. O. Edvardsen, A.L. Reiersen, M.W. Beukers, K. Kristiansen, tGRAP, the G-protein coupled receptors mutant database, Nucleic Acids Res., 30 (2002) 361–363.
32. D. Hanisch, K. Fundel, H.-T. Mevissen, R. Zimmer, J. Fluck, ProMiner: rule-based protein and gene entity recognition, BMC Bioinformatics, 6 Suppl 1 (2005) S14.
33. A.K. McCallum, Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering, 1996.
34. ISO 32000-1:2008, (n.d.).
35. E.A. Kabat, T.T. Wu, H. Bilofsky, M. Reid-Miller, H. Perry, Sequence of Proteins of Immunological Interest, 1983.
36. B. Al-Lazikani, Standard conformations for the canonical structures of immunoglobulins, Journal of Molecular Biology, 273 (1997) 927–948.
37. K.R. Abhinandan, A.C.R. Martin, Analysis and improvements to Kabat and structurally correct numbering of antibody variable domains, Mol. Immunol., 45 (2008) 3832–3839.

38. W.W. Chapman, P.M. Nadkarni, L. Hirschman, L.W. D'Avolio, G.K. Savova, O. Uzuner, Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions, Journal of the American Medical Informatics Association, 18 (2011) 540–543.
39. Access denied?, Nature, 462 (2009) 252–252.
40. S.C. Garman, Structure-function relationships in alpha-galactosidase A, Acta Paediatr Suppl, 96 (2007) 6–16.
41. R. Grantham, Amino acid difference formula to help explain protein evolution, Science, 185 (1974) 862–864.
42. S. Henikoff, J.G. Henikoff, Amino acid substitution matrices from protein blocks, Proc. Natl. Acad. Sci. U.S.A, 89 (1992) 10915–10919.
43. J. Amberger, C.A. Bocchini, A.F. Scott, A. Hamosh, McKusick's Online Mendelian Inheritance in Man (OMIM), Nucleic Acids Res, 37 (2009) D793–796.
44. S.T. Sherry, M.H. Ward, M. Kholodov, J. Baker, L. Phan, E.M. Smigielski, K. Sirotkin, dbSNP: the NCBI database of genetic variation, Nucleic Acids Res, 29 (2001) 308–311.
45. D. Fredman, M. Siegfried, Y.P. Yuan, P. Bork, H. Lehväslaiho, A.J. Brookes, HGVbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources, Nucleic Acids Res., 30 (2002) 387–391.
46. P.D. Stenson, E.V. Ball, K. Howells, A.D. Phillips, M. Mort, D.N. Cooper, The Human Gene Mutation Database: providing a comprehensive central mutation database for molecular diagnostics and personalized genomics, Hum. Genomics, 4 (2009) 69–72.
47. Y.L. Yip, H. Scheib, A.V. Diemand, A. Gattiker, L.M. Famiglietti, E. Gasteiger, A. Bairoch, The Swiss-Prot variant page and the ModSNP database: a resource for sequence and structure information on human protein variants, Hum. Mutat, 23 (2004) 464–470.
48. P. D'Ursi, F. Marino, A. Caprera, L. Milanesi, E.M. Faioni, E. Rovida, ProCMD: a database and 3D web resource for protein C mutants, BMC Bioinformatics, 8 Suppl 1 (2007) S11.
49. H. Venselaar, T.A.H. Te Beek, R.K.P. Kuipers, M.L. Hekkelman, G. Vriend, Protein structure analysis of mutations causing inheritable diseases An e-Science approach with life scientist friendly interfaces, BMC Bioinformatics, 11 (2010) 548.
50. R. Kanagasabai, K.H. Choo, S. Ranganathan, C.J.O. Baker, A workflow for mutation extraction and structure annotation, J Bioinform Comput Biol, 5 (2007) 1319–1337.
51. P. Kumar, S. Henikoff, P.C. Ng, Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm, Nat Protoc, 4 (2009) 1073–1081.
52. I.A. Adzhubei, S. Schmidt, L. Peshkin, V.E. Ramensky, A. Gerasimova, P. Bork, A.S. Kondrashov, S.R. Sunyaev, A method and server for predicting damaging missense mutations, Nat. Methods, 7 (2010) 248–249.
53. Interactive Biosoftware, Alamut, Interactive Biosoftware, n.d.

3DM: systematic analysis of heterogeneous super-family data to discover protein functionalities

Remko Kuipers, Henk-Jan Joosten, Willem van Berkel,
Nicole Leferink, Erik Rooijen, Erik Ittmann,
Frank van Zimmeren, Helge Jochens, Uwe Bornscheuer,
Gert Vriend, Vitor Martins dos Santos, Peter Schaap

Abstract

Ten years of experience with Molecular Class Specific Information Systems (MCSIS) such as with the hand-curated G protein-coupled receptor database (GPCRDB), or the semi-automatically generated nuclear receptor database (NRDB) has made clear that a wide variety of questions can be answered when protein related data from many different origins can be flexibly combined. MCSISes revolve around a multiple sequence alignment that includes "all" available sequences from the entire super-family, and it has been shown at many occasions that the quality of these alignments is the most crucial aspect of the MCSIS approach.

We describe here a system, called 3DM that can automatically build an entire MCSIS. 3DM bases the multiple sequence alignment on a multiple structure alignment, which implies that the availability of a large number of super-family members with a known three-dimensional structure is a requirement for 3DM to succeed well.

Thirteen MCSISes were constructed and placed on the Internet for examination. These systems have been instrumental in a large series of research projects related to enzyme activity or the understanding and engineering of specificity, protein stability engineering, DNA-diagnostics, drug design, etcetera.

Introduction

Due to new high-throughput DNA sequencing, protein crystallization technologies, and efficient large scale mutant generation techniques such as chemical gene synthesis, there is an exponential growth of the amount of data submitted to databases such as the PDB [1], UniProt [2] and PubMed [3]. Data deposition rates are getting so high that screening new database depositions just for molecules of one family of interest already requires specialized software.

Each protein related data type can be a powerful information source in itself, but combined, they can become true eye-openers. A good example is probably the combination of protein sequences and structures into structure-based multiple sequence alignments. These alignments contain a wealth of evolutionary fingerprints such as correlated mutations or amino acid conservation and variability patterns that can be used to get clues about the function of individual amino acids. Alignments can be used in combination with many other, often heterogeneous, protein data types. Mutation information, for instance, can be successfully transferred between proteins using an alignment. A mutation that influences antagonist binding in one nuclear hormone receptor (NR), for example, is likely to have a similar effect in other NRs. Another example is the combination of keywords in PDB files with the presence of specific amino acids at specific positions in an alignment. This latter concept was published in 2004 when Folkertsma et al. used this approach to detect a residue responsible for antagonist binding in the NR super-family.

The above examples show that, when studying individual members of a protein family, much can be gained when all available information about all individual members of the super-family is available concurrently. This is often difficult to achieve because the processes of data collection, validation, storage, and the subsequent determination of correlations between heterogeneous data types can be technically complicated, laborious, and very time consuming. The data collection and storage process suffers from a large number of technical complications such as the existence of protein isoforms, or different numbering schemes used for homologous proteins. Other data available at the amino acids level such as mutation data, ligand binding data, etc., can also have their own nomenclatures and numbering schemes. Especially the residue numbering problem must be solved rigorously before correlations between the various data types can be studied.

For an easy comparison of heterogeneous super-family data several web-interfaced Molecular-Class-Specific Information Systems (MCSIS) have been developed that store, combine, validate, give structure, and present data about protein super-families. The GPCRDB [4] is a MCSIS for G protein-coupled receptors. This system, originally designed in 1993, has been used as a model system for the nuclear hormone receptors (NuclearDB) [4], potassium channels (KchanelDB) and prions (PrionDB) MCSIS. An important lesson learned from the design and use of these MCSISes was that the quality of the MSA is the most important aspect of any subsequent research performed, and indeed, the curators of these

MCSISes have spent many months or even years on obtaining the best possible alignments. In this process new profile-based sequence alignment methods were designed that would allow for user interaction; mutation collections were made and used to obtain information about particularly hard to align sequence sections, ligand binding studies were used to reveal whether residues in different proteins were, or were not likely to be at equivalent locations in their structures. These systems thus relied on extensive manual curation, which makes it difficult to keep them up-to-date. Nevertheless, the power of having so much heterogeneous data at one's disposal in one coherent system allows for a wide variety of theoretical and applied scientific questions to be addressed [5-18].

The 3DM system provides a generic framework to exploit heterogeneous protein related data, and facilitates a broad range of tasks needed to automatically create and update the next generation of MCSISes. It generates a highly accurate multiple sequence alignment (MSA) of a protein super-family by using a multiple structure alignment (3D-MSA) to guide the construction of the (much larger) MSA. From the 3D-MSA it generates a unified 3D numbering scheme, which is subsequently used for all protein related data types including aligned structures and sequences, homology models, and mutational and computationally derived data. This unified numbering also enables the easy construction of homology models. Most data types are automatically mapped onto the alignment and can also be visualized in the 3D-models and structures of individual proteins.

Here we present the 3DM framework that is at the basis of a series of MCSIS systems and illustrate the possibilities with a series of highly successful experiments performed in the fields of protein engineering, drug design, and DNA diagnostics.

Materials & Methods

Definitions

Within a 3DM system a super-family is defined as a group of structures with a common structural fold. A sub-family is the smallest building block of the alignment and is a profile based alignment starting from a single structure. Families are groups of related sub-families.

Data Collection

Data collection (Fig 1. Step 1) is performed by a series of scripts – one script per data source. Scripts have been written to extract information from Swiss-Prot (http://www.expasy.org/sprot/), NCBI (http://www.ncbi.nlm.nih.gov/), literature, etc. In addition, the flexible set up of 3DM facilitates easy extraction of data from specific sources. For example, 3DM contains scripts to extract data from two public p53 mutation databases (http://www-p53.iarc.fr/, http://p53.free.fr/).

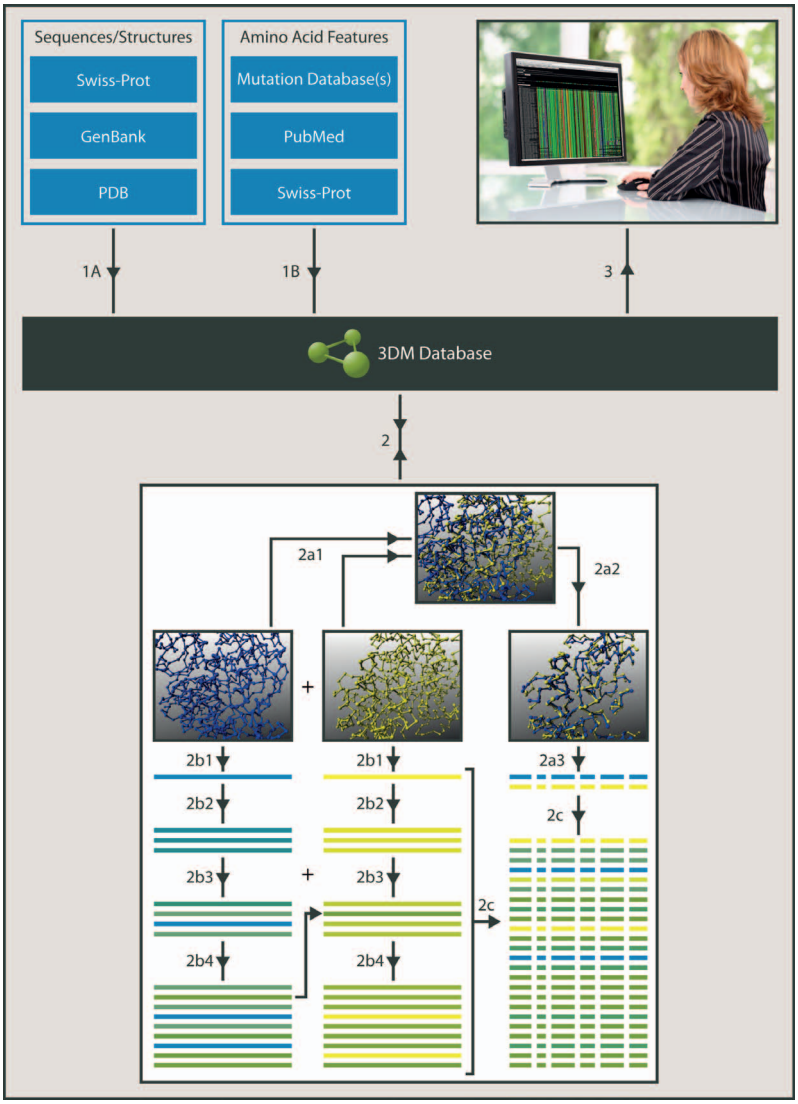


Figure 1. The 3DM generic framework at a glance. Step 1 Data collection. 1A: Sequences and structures are automatically inserted into the 3DM database. 1B: Features related to amino acids, such as mutational information from the literature, are collected, and automatically inserted in the database. Step 2 Creation of the 3D-MSA. Step 2a1 Superposing of the structures. Step 2a2 Deletion of structural variable regions and (2a3) core determination. Step 2b illustrates the iterative profile based alignment procedure to create the individual subfamily alignments in four (default value) successive alignment rounds. In step 2c the separate subfamilies are aligned to one large super-family alignment using core data. (step 2a for details see materials and methods). Note that some individual sequences can be confidently aligned to different structures here indicated by the arrow showing a sequence aligned in both subfamily alignments. These sequences are used for alignment quality control (see materials and methods). Correlated data can be visualized by the user via HTML pages (step 3)

Structural information was obtained by extracting family members from SCOP [19] and running BLAST [20] against the PDB database using as queries representative members from each SCOP subfamily. Representative structures show less than 80% (3DM default cut-off) pair-wise sequence identity. BLAST searches were also performed with these same sequences to extract all super-family sequences from the Swiss-Prot and Trembl databases.

Multiple Sequence alignment

3DM uses a three-step procedure (Fig. 1, step 2) to create a super-family alignment. In the first step (Fig. 1, step 2a1), all collected structures of a super-family are superposed [21] and a common core of structurally equivalent positions (called core positions) is determined (Fig. 1, step 2a2). The WHAT IF structure superposition method was described already in 1991 [21], and will here be described very briefly. WHAT IF collects from both structures all fragments of 15 amino acids or longer and does an all-against-all comparison on these fragments. This is done in distance space rather than Cartesian space, which makes this process very fast. Each pair of matching fragments between the two structures is related via a superposition matrix, and when two pairs of matrices have a similar trace, it is highly likely that the two pairs of fragments can be superposed together. This is tried in Cartesian space. This process is fast enough so that it can be repeated for all possible clusters of pairs of fragments, and the 10 largest clusters are refined using an optimized version of the algorithm of Rossmann and Rao [22]. As the Rao and Rossmann method is known to occasionally find the wrong local minimum, we restart the best cluster 25 times; each time using a global superposition matrix that is randomly rotated 5-15 degrees in a random direction. The quality of this method has be described in an article in which a series of methods was compared [23]. The conclusion was that the WHAT IF method is especially suitable for the accurate detection of superpositions of large numbers of non-contiguous fragments that fit well. This is the main reason why this method was selected as Oliveira *et al.* found that the quality of alignments is crucial for a series of follow-up calculations such as correlated mutation analyses [12,13].

The core is typically defined as the collection of residue positions where at least three consecutive residues have their alpha carbon located within 2.5 Ångstrøm of the average position in at least 90% of the superposed structures. Cores can differ depending on the order in which the structures are superposed. For example, starting with structure A, a structurally conserved helix might be detected in structure B, but, even when present in structure C, this helix might just fall outside the cut-off. Starting with structure B, however, the helix in C might be detected and then it is missed in A. If in B it is positioned in-between the helices of A and C, then the structure alignment order A, C, B will allow us to find all three helices as structure mates. 3DM therefore repeats step 2a starting with one representative structure of each distinct subfamily. For completeness all obtained core files are then merged into one super-family core.

In the second step (Fig. 1, 2b), separate subfamily alignments are created by aligning to each of the representative structures all sequences with more than typically 30% (3DM default value) sequence identity. The alignments are created by WHAT IF [24] using the sequence-to-structure iterative profile based alignment procedure as described by Oliveira *et al.* [25].

In the third step (Fig. 1, 2c), the resulting single-structure-based subfamily alignments are merged into one large alignment guided by the structure alignments from step one. Each subfamily alignment comprises all residue positions present in its parent structure, whereas the combined alignment only contains 'core' residue positions.

Alignment quality

In the 3DM approach, correctness of the alignment is more important than completeness. Therefore, 3DM uses several alignment quality checks before a sequence is included in the alignment (Fig. 1 step 2c). All sequences that can be aligned to one of the representative structures with a sequence identity above a threshold (inclusion-cut-off; present 3DM default value = 55%) are automatically included in the alignment. If a sequence is aligned against more than one structure with the inclusion-cut-off above the threshold, then the alignment with the highest sequence identity is used. We have observed that mapping of sequences onto the core is independent of the choice of structure against which a sequence is aligned, as long as the inclusion-cut-off of 55% is met. In our experience small cores are more difficult to align and demand a higher inclusion-cut-off.

Sequences that show upon alignment a percentage identity between the lower cut-off (present 3DM default value = 30%) and the inclusion-cut-off are only included if they can be aligned against two representative structures and if both alignments are consistent in terms of core assignments. If a sequence is aligned against three or more structures the most abundant core assignment is used.

Parameter choices

3DM provides a series of user-adjustable parameters like sequence identity cut-off, percentage of structures that must participate at any residue position in the superposition for acceptance, gap open and elongation penalties for each of the alignment procedures, minimally required sequence identity percentages, etc. Plausible default values were selected for each parameter based on our experience with building a dozen 3DM systems. Systematic optimization of these parameters is still beyond the realm of possibilities of today's computers.

3DM -numbering scheme

A semi-arbitrary sequence numbering scheme is applied in which residues are numbered sequentially starting at 1 for the most N-terminal core position. These numbers are referred to as 3D-numbers. These 3D-numbers are used throughout all alignments and are incorporated in all structure files and homology models.

Correlated mutations

3DM contains a newly developed tool, Comulator, for correlated mutation analyses (CMA), of large structure-based multiple sequence alignments. The tool and results obtained by Comulator are published elsewhere [9]. The CMA scores that result from Comulator are stored in the 3DM database connected to the 3D-numbers. This makes the CMA scores easy to relate to the other data types. The correlations are visualized in hyperlinked heatmaps enabling the selection of a subset of sequences that contain a certain amino acid couple at a correlating position pair. These subsets can be used to create a filter (see the section on filters) from which a new 3DM (sub)system can be generated automatically.

Ligand contacts

3DM uses WHAT IF for the extraction of contact data between co-crystallized compounds (ligands) and protein molecules from all super-family PDB files using a method similar to ARES [26]. The 3DM definition of a ligand is a molecule with more than 6 atoms that is not a solvent molecule commonly used in protein crystallization (e.g. sulphate, glycerol, etc). All contacts are stored in the 3DM database connected to the 3D-numbers. They are visualized in the context of other data in the alignment detail pages and in the protein structures.

Mutations and other amino acid features

3DM contains a literature mutation extraction tool, Mutator, which is modelled after MuteXt [27]. Mutator uses the protein descriptions from a super-family as keyword searches in the PubMed database and extracts all mutations from the resulting literature. Mutator and results obtained by Mutator are published elsewhere [28]. Swiss-Prot holds many data about protein features such as post-translational modifications, mutations, and natural variants, metal binding, etc. 3DM collects and stores these features. The p53 MCSIS (manuscript in preparation), for instance, illustrates the flexibility of 3DM because additional mutation information could be collected from two large mutation databases (the IARC TP53 Mutation Database[29] and the UMD-p53 database [30]) and simply incorporated and used in the MCSIS with all other (mutation) data.

Filters and subsystems

3DM contains several filter methods to obtain a sub-selection of proteins from the 3DM database. This filtering system enables the automatically generation of a new MCSIS for a subset of sequences. This subset system remains fully connected to the complete superfamily MCSIS. All 3DM features, such as alignment, correlated mutations, alignment detail graphs, visualisation in structures, etc., are also available for a subset MCSIS. 3DM provides multiple filters to select a subset of the proteins. For instance, a subset can be generated that contains sequences only from just one or more subfamilies. A subset can be retrieved using BLAST searches, sequence motifs, selection of a sub-branch of the phylogenetic tree, by keyword searches in PDB files, protein names, organism names, etc. Separate subsets of sequences that result from different selection options can be logically combined. The search filters of 3DM were used, for example, to query structure files in the nuclear receptor 3DM in order to generate two different subsets of sequences related to a) PDB files with an agonist in the ligand binding domain and b) PDB files with an antagonist.

3DM Interface

Usage of a 3DM database is organized via interactive web pages that interact with the underlying database (Fig. 1, step 3). The homepage of 3DM is directly linked to the collection of alignments. An alignment is the main point of access to the data underlying each 3DM, such as sequences, mutation data, 3D structures, etc., and to derived results such as CMA scores, 3D models, contact information, variability, and conservation in the alignment, etc. The first method for navigation through the 3DM information is provided from the alignments via links to four different types of secondary pages:

1. Protein detail pages, which can be retrieved via the protein names. The protein detail pages provide information about the similarity of the sequence aligned against different structures, the 3D models (superposed on the 'core' and numbered as desired), raw sequence data, and a collection of all mutations that 3DM has collected for the protein.

2. Amino acid detail pages. These are linked to the amino acids in the main alignment. The amino acid detail pages provide amino acid features, such as mutational data and the corresponding links to the literature. Additionally, structure models can be retrieved from these amino acid detail pages in which computationally modelled mutations have been introduced. These structure models, which contain the 3D-numbers, are visualized with a sphere around the in silico introduced amino acid to indicate possible bumps with surrounding amino acids.

3. Position detail pages. These can be retrieved via the consensus sequence on top of the alignment. The position detail pages provide CMA results, amino acid distribution histograms, integrated positional mutation information from all sequences in the alignment, integrated contact information from all structure files, etc.

4. (Sub)-family detail pages. These can be retrieved from the scroll down button at the top of each sub-family. The family detail pages are separate 3DM systems within the super-family 3DM. A family contains a selection of sub-families that are evolutionary close together. A 3DM family is similar to a filter with one important difference: In contrast to a filter the core of a 3DM family contains more residues than the core of the super-family. The sub-family detail pages provide an alignment of full length sequences. The alignment of the non-core regions of the sequences is done by Muscle [31]. At the subfamily detail pages mutational information is also visualized in these regions.

The second method of navigation through the 3DM data is via the scroll down menus which are present in all 3DM web pages. The 'Families' scroll down menu is hyperlinked to the available 3DM (super)-families. The 'Alignment' menu provides links to the various different alignment views, the correlation data, a page to insert data in the 3DM database (only for administrator users), and alignment statistics for data comparison (see section comparison of data types). With the 'Filter' option filters can be generated and updated (see section filters). The 'Search' option provides various options to query the 3DM data and to generate a sub-selection of sequences which can be used to generate a filter. The 'View' option provides links to the alignment detail pages, the correlation data, a protein structure detail page, a taxonomy page, and to pages for data visualisation in protein structures.

Homology models

3DM uses YASARA (www.yasara.org) to visualize the super-family protein structures and 3DM homology models. The 'VIEW->Data in Structure' option of 3DM enables the user to select multiple protein structures (these are all superimposed), multiple co-crystallized compounds (even from different protein structures than those selected) and different data types, such as correlated mutations, contact information, charge clamps, conservation, etc. All selected data will be written in a so-called YASARA scene file that can be read by the free YASARA viewer available from www.yasara.org. All selected homology models use the 3D-numbering scheme, which make it easy to compare them with the 3DM alignment and the underlying data.

Availability

The 3DM systems discussed here are available at http://3dmcsis.systemsbiology.nl and http://3dmcsis.systemsbiology.nl/FMDB/ The alignment-related software, especially the 3D-MSA part is performed using the WHAT IF software (free for academic use). WHATIF macro's used to produce an example MSA are available from http://3dmcsis.systemsbiology.nl/software/

Results and Discussion

Implementation

An introduction into the technology behind 3DM is given in fig. 1. Sequences and structures of a target super-family are automatically extracted from the public repositories and inserted into the 3DM database. Input structures are used to create a superposed coordinate file. An iterative profile-based sequence alignment procedure is used to create the individual subfamily alignments. The separate subfamily alignments are aligned guided by the superposed coordinate file. Features related to individual amino acids of individual

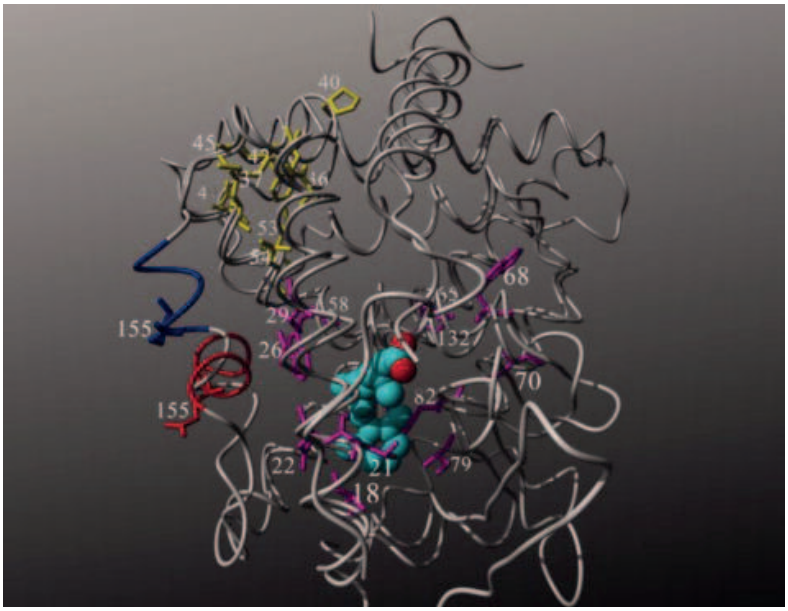


Figure 2. Superposing of the agonistic (pdb code 1XDK) and antagonistic (pdb code 1L2J) of the ligand binding domain of a nuclear receptor. A co-crystallized ligand is shown in ball representation. Residues at positions with high CMA scores are purple (3D-numbering scheme) Most of these residues are located at the rim of the ligand binding pocket. Highly conserved residues (>85%, yellow) are located at the co-factor binding site. In the agonistic conformation is helix 12 close to the ligand binding pocket and the conserved glutamate 155 is positioned to bind the cofactor for activation (here depicted in red). The positions of helix 12 and its glutamate in antagonistic conformation are in blue.

sequences, such as mutational information are also automatically extracted from the literature. It has proven difficult to quantify the quality and completeness of automatically extracted data. The collection of structure files, for example, is complicated because none of the currently available tools for obtaining similar 3D structures (e.g., SCOP, Structure Matching tool [32], BLAST, etc.) provides a complete collection. Some of these systems look only in non-redundant subsets of the PDB, others have algorithmic features that exclude certain structures, and yet others are not based on the latest version of the PDB. Therefore several structure extraction methods were implemented and used in parallel to obtain as many structure files as possible, even though this could result in duplicated structure information, which must be removed later. Superposing is done automatically using WHAT IF software. Unfortunately sometimes many highly similar superposition solutions exist and not always all structure files can be used because they contain too many syntactic errors for automatic superposing. The collection of primary sequences, on the other hand, is much simpler because software like BLAST can easily be parameterized to include every possible homolog. In the 3DM approach, correctness of the 3D-MSA is much more important than completeness. Some remote homologs therefore are not (yet) included because the structures required for a reliable alignment are not available.

The definition of the core set of structurally superposed residue positions (fig. 1 step IIa) depends to some degree on the cut-off parameters chosen by the operator. The 3DM automatic superposition and alignment module was benchmarked against a manually created structure-based multiple sequence alignment of the nuclear receptors [26]. The manually created NR alignment resulted in a 'core set' of 183 positions. 3DM automatically detected a core of 158 residues. Most of the core positions that were not automatically detected are involved in movement of helix I and helix XII (fig. 2). These residues were manually added to the core by Folkertsma *et al.* even though they did not superpose at all. 3DM does allow the curator to manually extend the core between steps IIa and IIb (fig. 1). In 2004 Folkertsma *et al.* collected 1,577 NR sequences, 468 of which they could reliably align. Using the same set of input sequences, 3DM could automatically align 752 sequences. Although this increase is mainly a result of the larger number of structures available it also demonstrates the need for automation. In just two years the number of "alignable" NR sequences in the 3D-MSA increased by 37%. This increase in number of alignable sequences resulted in a concomitant increase in usable information. For example, in 2004, no mutation information was available for 38 of the 158 core positions, whereas currently mutation information is available for 150 core positions. In nine cases, this gain of information was due to the increased number of aligned sequences. In 21 cases this was due to new mutational information automatically extracted from the literature. For instance in 2004, no mutation information was available for a conserved alanine in the WAK motif in helix III and contradicting predictions were made for the role of this residue. A recently published mutation of this alanine to a lysine in the androgen receptor [33] was automatically extracted from the literature. As the high preference for an alanine residue at the particular residue position in the 3D-MSA already suggested,

this mutation had an effect on ligand binding. The last update of the NR (February 2009) contains 2152 sequences. From the literature 3DM automatically extracted 2,539 mutations covering 98% of the structural core.

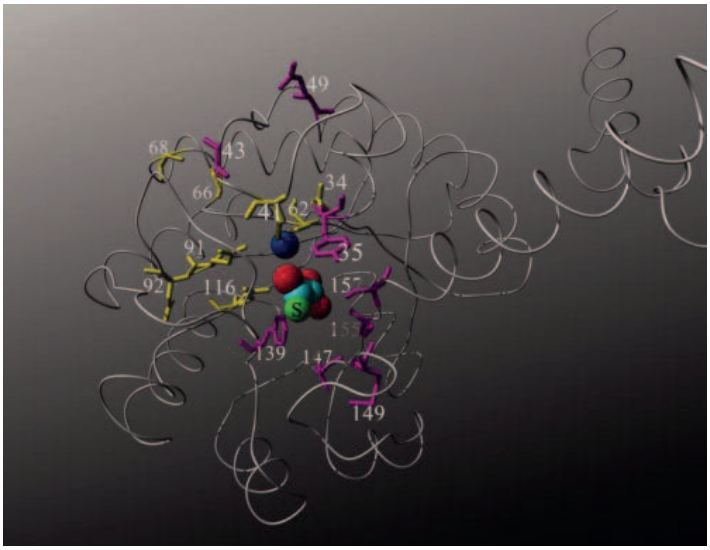


Figure 3. Conserved and highly correlated positions in the model structure of oxaloacetate hydrolase from A. niger. The 3D-numbering system is used. Residues at conserved alignment positions (>98%) of the PEPM/ICL super-family are yellow. Most of these residues are located near the conserved oxalate moiety of the substrate and the conserved metal ion (blue). Residues at alignment positions having a high CMA score are in purple. Most of these residues form a network located near the substrates substituent group (S) in green.

Limitations

Today, structure information can be found for at least one detectable homolog for about 40% of all known sequences. Obviously, 3DM can work without structures, but in that case the multiple sequence alignment will be no better than the HSSP-based profile alignment method that we use for multiple sequence alignment purposes, and all advantages of the structure based approach will be lost. Correlated mutations analyses tend to work best when the alignment contains hundreds till thousands of sequences [12,13]. However, if the signal is clear, i.e., the alignment is correct and the spread in the sequences is adequate, then even a small number of sequences can already reveal useful information. As an example we made a 3DM for the very small family of crambin-like proteins. The PDB database holds some 30 structures of crambin variants while the total number of unique sequences that can be extracted from UniProt is currently about 50. Most crambin-like molecules have three conserved cysteine bridges. The highest 3DM CMA score is between residue positions 12 and 29, positions which in a few of the family members form a fourth cysteine bridge.

Use Cases

3DM was initially designed for the purpose of guiding protein engineering projects. Protein engineering projects are usually focussed on increasing enzyme activity or on changing or modifying enzyme specificity. Despite many computational efforts, it still proves difficult to consistently predict the specific mutations needed to achieve such goals. Most protein engineering project therefore still rely on random mutagenesis approaches, where thousands or even hundreds of thousands mutants are generated and screened for the desired effects. Nowadays, with chemical techniques for gene synthesis, mutant libraries can be generated with combinational changes at only a few specific positions in the protein. Narrowing down the library size to specific positions makes subsequent screening much more efficient. Correct selection of these protein engineering hot spots and selection of the amino acid residues that should be introduced at these positions is then crucial. Below we demonstrate how the 3DM framework was used to predict such hot spots for protein specificity, protein activity and for thermostability.

Substrate Specificity

The PEPM/ICL super-family consists of a variety of enzymes that can break carbon-carbon bonds, such as lyases (isocitrate lyase), hydrolases (oxaloacetate hydrolase OAH), and mutases (phosphoenolpyruvate mutase). All these enzymes act on a bond between a carbon and an oxalate-like moiety that is always bound to an Mg^{2+} or Mn^{2+} metal ion in the active-site cleft (fig. 3). 3DM detected a network of nine positions 43, 49, 75, 139, 147, 149, 150, 155, and 157 (3D-numbering) with high CMA scores located in the active-site cleft in the area where the substituent group of the substrate binds (fig. 3). This strongly suggested that this part of the active site is involved in substrate specificity. To test this hypothesis we have mutated one of these positions in OAH of Botrytis cinerea [34]. Position 157, a serine in OAH, was selected since this position is closest to the substituent group of the substrate (Fig. 3). Three mutations were made: S157T, S157P, and S157A, since these three residues are the most abundant residues at position 157 (42%, 38%, and 11% respectively). As expected, these mutations did not significantly decrease activity of the protein, but they did have a drastic effects on the affinity of OAH for its substrate oxaloacetate [34]. This suggested that this position could be considered as a hot spot for changing substrate specificity. To test this hypothesis, mutations were made at the same position in two structurally related proteins. The first protein that was mutated was the petal death protein (PDP) which can convert multiple substrates including oxaloacetate. Like OAH, PDP [35] has a serine at 3D number 157. The same S157T and S157P mutations were introduced in PDP, which specifically eliminated OAH activity from PDP while retaining all others [34].

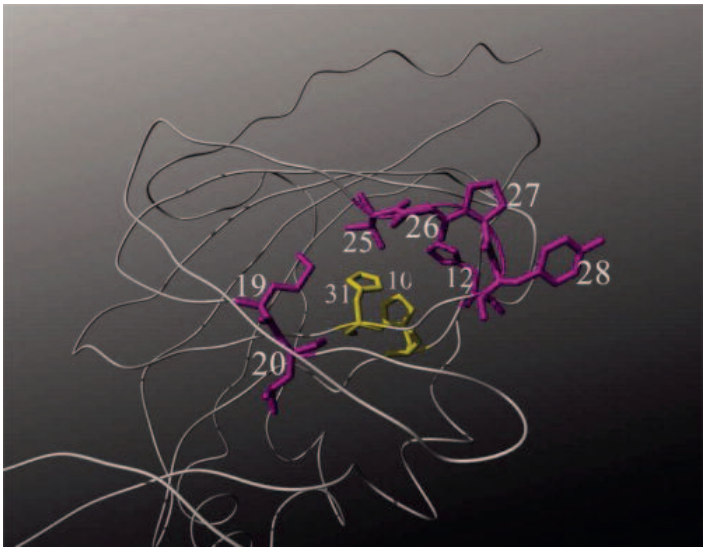


Figure 4. Tube representation of the 3D structure of PfPGI from P. furiosus (PDB accession code 1X82). The 3D-numbering scheme is used. The two conserved histidine residues are in yellow, the seven positions with high CM scores are in purple.

The second protein that was mutated at position 157 was methyl isocitrate lyase (MICL) which has a threonine at this position. In order to introduce OAH activity in MICL the complementary T157S mutation was made in MICL from E. coli (data not shown). Although this mutant could not convert oxaloacetate in oxalate and acetate, the T157S mutation, again had a drastic effect on the affinity of MICL for its natural substrate whereas the activity was virtually unchanged. This example shows that mutational information can successfully be transferred between different proteins, even though these proteins belong to different sub-families. From the fact that mutating position 157 in different subfamilies leads to the same effect we can conclude that this position is correctly aligned. This is especially rewarding because the region around position 157 is sequentially highly variable.

To demonstrate the connectivity of heterogeneous data types in a 3DM and to show the usability of the 3DM filtering system, a subset of OAH-like proteins from known oxalate producing fungi was generated. For each filter a graph is automatically generated that highlight residues specifically conserved in the corresponding subset of sequences. These "filter specific conservation" graphs can be retrieved from the position detail pages (see section 3DM interface of M&M). Not surprisingly, the OAH specific serine was the number one residue in this "filter specific conservation" graph of oxalate producing fungi. This example demonstrates the power of combining heterogeneous data. The multiple sequence alignment, the CMA, together with the possibility of rapidly determining the correlation between a physiological trait, residue type, and enzyme class brought position 157 forcefully to our attention.

Enzyme activity I

The proteins that belong to the RmlC-like cupin super-family are functionally diverse [36], but most are enzymes of which the active site is located within the β-barrel. This active site often contains two histidines on core positions 10 and 31 that are conserved in approximately 80% of all 2,035 sequences (fig. 4: magenta residues). The 3DM Comulator tool revealed a network of highly correlating positions (3D-numbers 12, 19, 25, 26, 27 and 28), of which the highest pair-wise CM score was detected for position pair 27 and 28. This pair is located in a structurally conserved surface loop far away from the active site histidines (fig. 4). We have studied how the relationship of the two positions affects enzyme activity in Pyrococcus furiosus PGI (PfPGI) [9]. PfPGI has a tyrosine at position 28 and first we introduced a glycine since, with 42%, this is the most prevalent amino acid at this position in the cupin super-family. The Y28G mutant (residue Y133G in PfPGI sequence) resulted in a 2.6 fold reduction of the activity. The most abundant residue at position 27 is also a glycine which is present in 26% of the aligned sequences. PfPGI however has a proline at this position. Analysis of the Comulator heat-map reveals that the PfPGI combination P27-Y28 occurs rarely (1.1%), whereas P27-G28 occurs in 4.4% of all sequences. However, when a glycine is present at position 28, by far the most prevalent residue at position 27 is an alanine (23%) followed by a glutamate (18%) and an arginine (18%). The double mutant P27A/Y28G not only regained the activity lost by the Y28G mutant, but even became twice as active as the wild-type enzyme. Also both the P27E/Y28G and P27R/Y28G double mutants could compensate for the loss of activity of Y28G, regaining near WT activity. It should be noted that in practice these results can not be obtained with a random mutagenesis and screening approach.

Enzyme activity II

Members of the vanillyl-alcohol oxidase (VAO) flavoprotein family catalyze a wide variety of oxidation-reduction reactions and share a characteristic domain topology, with a conserved N-terminal FAD-binding domain and a less well conserved C-terminal cap domain. Most VAO-family members are oxidases that bind the FAD cofactor in a covalent manner [37]. To rationalize these properties with only few protein structures available, a 3DM alignment of the N-terminal domain using 1152 sequences of (putative) VAO-family members was constructed. It turned out that position 36 (a histidine in members with a covalently bound flavin) correlates well with positions 78, 91 and 92[9]. Glycines are favoured at positions 78, 91 and 92, and they are all located in the vicinity of the FAD. L-galactono-g-lactone dehydrogenase (GALDH), a VAO-family member lacking oxidase activity, contains a leucine at position 36 (Leu56), an alanine at position 78 (Ala113), and a valine at position 91 (Val126). Ala113 was identified as a potential key residue, as it is predicted to be near the reactive flavin-C4a position creating a steric block that prevents oxygen from approaching the enzyme active site. Mutation of Ala113 to Gly generated an enzyme with 400-fold higher oxygen reactivity than wild-type GALDH, similar to that of other VAO family

members [38]. In the wild-type enzyme Ala113 acts as a gatekeeper, preventing oxygen to access the isoalloxazine nucleus. The pr esence of such an oxygen access gate seems to be a key factor for the prevention of oxidase activity within the VAO family, and is absent in members that act as oxidases. This study together with results from molecular dynamics simulations [39] provides an important advance in our understanding of the selective reactivity of flavoproteins toward oxygen.

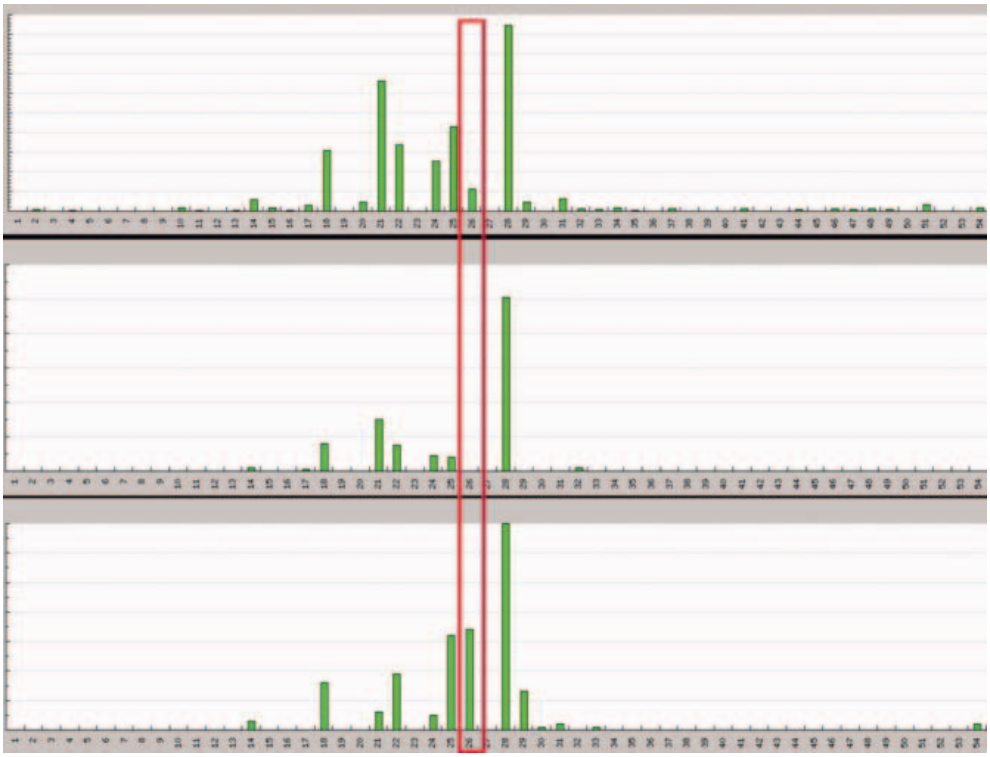


Figure 5. 3DM contact data plots from the nuclear receptor ligand binding domain. Horizontal axes; alignment positions (3D-numbering scheme). The green bars indicate the number of contacts per position retrieved from PDB structure files. Upper panel; number of contacts retrieved from all relevant PDB files. Middle panel; contacts retrieved from PDB files with an agonistic compound. Lower panel; contacts retrieved from PDB files with an antagonistic compound. Alignment position 26 (boxed) makes only contacts with antagonistic compounds.

Thermostability

A 3DM alignment consisting of 2,813 α/β-hydrolase super-family members including esterases, dehalogenases and epoxide hydrolases was used to increase the thermostability of an esterase from Pseudomonas fluorescens. Three positions were determined to be effective towards thermostability using B-FITTER (http://www.mpi-muelheim.mpg.de/kofo/institut/arbeitsbereiche/reetz/reetz_e.html). From the determination of the amino acid

distribution at these three positions using the 3DM software, codons were elucidated that incorporate only those residues that are frequently present in the natural set of enzymes. By using this strategy, the library size could be substantially reduced from ~100,000 to only 3,450 variants to cover 95% of all possible combinations in a simultaneous saturation experiment. After preparation and screening of this library, the best mutant showed a melting point 8°C higher than the wild-type enzyme. To further support the quality of the mutant library, two additional libraries were created in which either all 20 natural amino acids or only those that appear very rarely at these positions were incorporated and representative numbers of variants were screened for activity and thermostability. The comparison of all three libraries concerning the hit rate and the average quality of hits (higher thermostability while maintaining wild-type activity) clearly showed that beside the advantage of less screening effort, the quality of the smaller, 3DM driven library, is much better compared to the incorporation of all 20 amino acids and even more if only seldom residues are introduced (Jochens, H., Bornscheuer, U.T., manuscript in preparation).

3DM in drug design

Drug design is a very elaborate project, and when a new drug hits the market one can be almost certain that thousands of scientists have thrown in the skills of their specialism to come to the final result. The early stages of a drug design pipeline are characterized by a massive collection and generation of heterogeneous data. Sequence – structure – function relation studies are a crucial early step necessary to obtain the information needed to guide, for example, small molecule library design, lead optimisation, or toxicogenomics studies.

Folkerstma *et al.* were especially interested in in silico discrimination of agonistic and antagonistic small molecules [8]. They manually created an MCSIS for the nuclear receptor (NR) ligand binding domain, and described a correlation method that allowed them to find residues that explain the binding and mechanistic differences of agonists and antagonists. Their study was quite laborious because it involved; i) manual generation of a high quality structure-based multiple sequence alignment; ii) extraction from the structure files of contact data for agonistic and antagonistic ligands; iii) correlating contact data, etc., with the sequences in the alignment. It took Folkertsma et al., more then a year to make the MCSIS and to do the analyses. The 3DM NR required one month, of which most time was simply a computer crunching through the thousands of sequences, structures, articles with mutation data, etc. Correlation studies like the one performed by Folkertsma et al., today only take minutes to complete, since all different data types are fully connected in the 3DM system and generation of the graphs (fig. 5) that form the bases of such studies is fully automated. Actually, the only limit on the possible correlation studies lies in our ability to think of new hypotheses when thinking about what can be done with the available data.

DNA diagnostics

The 1,000-dollar genome [40] will be generating more problems than it solves. Even if only the coding regions are sequenced of the genome of a person with a genetic disorder we can still expect thousands of deviations from the generic, 'wild-type' human genome. Such studies will have to be combined with powerful computational methods that can prioritize the observed differences in terms of likelihood of being involved in that disorder. The role of P53 in cancer is an example. P53s are transcription factors involved in tumour repression. P53 mutations are found in more than 50% of all cancer patients [41] and in 580 different kinds of tumours [42]. However, if among the thousands of mutations observed in the sequenced genome of a patient a mutation is detected in P53, we can still only make a strong prediction if that particular mutation has been proven correlated to cancer in previous studies. With over 60,000 observed mutations, the P53 3DM is the largest P53 mutation database available on the Internet. The phenotype (type of cancer) is known for almost all of these mutations, allowing for completely new areas of research. One striking observation is that 93.2% of the P53 cancer related mutations are observed in the 229 structurally conserved positions while only 6.8% are observed in the remaining 112 positions of the human P53 (results not shown).

The alpha-amylase 3DM contains more than 40 amylase sub-families, including the human alpha-galactosidase (hGLA). Fabry disease is a monogenetic X-linked lysosomal storage disease caused by mutations in hGLA. A deficiency in hGLA causes substrate accumulation, which ultimately leads to disease symptoms such as chronic pain, vascular degeneration and cardiac impairment. The diagnostics of Fabry (specifically the mild forms) can be difficult since the symptoms are similar for other diseases such as diabetics. The hGLA gene of a suspect patient is therefore sequenced to confirm the diagnosis. However, even if the gene indeed contains a non synonymous mutation, it still might be an uncommon non-pathogenic variant. In previous work (Kuipers *et al.*, manuscript in preparation) we have shown how the 3DM framework can assist in this type of DNA-diagnostics by correlating residue plasticity of core positions with specific mutations related to Fabry disease. The first major conclusion of this work is that 3DM could automatically collect more mutations from the literature then currently is stored in the conventional, mostly human curated, databases. 3DM collected 1,369 mutations (371 unique mutations) from the literature, whereas the human gene mutation database (HGMD [43]), which is the most complete database of Fabry related mutations, contains only 301 unique mutations. The second major conclusion was that there is a very strong correlation between the alignment derived information and a pathogenic Fabry mutations. First, like in P53, 75% of the reported Fabry mutations were associated with core positions. Second, there is a strong correlation between a disease related hGLA amino acid change and the occurrence of the changed residue at the corresponding alignment position in the alpha amylase 3D-MSA. For example, only 3% of the in total 1,034 reported hGLA mutations at core positions mutate to a residue that is present at the corresponding alignment position in more then 25% of the aligned alpha-amylase

sequences. Vice versa, 85% of all reported Fabry mutations located at core positions are mutated into amino acids which are present in the corresponding alignment position in less then 7% of the aligned alpha-amylase sequences. Clearly, introducing residues that are not frequently observed at the corresponding alignment position have a higher probability to be pathogenic.

Conclusions

The technologies of the omics era allow us to gather large volumes of data for many, often very different data types. These data must all be collected, validated, curated, and put in a common nomenclature and numbering scheme before they can be harvested through correlation techniques or other methods that have their basis in data integration.

The 3DM systems that were constructed so far have helped specialists in a series of research fields answer questions related to topics as diverse as protein stability, enzyme activity and specificity, the design of novel agonists or antagonists, and phenotype genotype correlations in humans with a genetic disorder.

These technologies will allow us to obtain and integrate all the time larger volumes of more data types. The subsequent deepening and widening of 3DM systems, in turn, will allow us to answer all the time more but also more complicated biological questions.

References

1. Dutta, S., Burkhardt, K., Young, J., Swaminathan, G. J., Matsuura, T., Henrick, K., Nakamura, H. & Berman, H. M. (2009). Data deposition and annotation at the worldwide protein data bank. Mol Biotechnol 42, 1-13.
2. Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N. & Yeh, L. S. (2005). The Universal Protein Resource (UniProt). Nucleic Acids Res 33, D154-9.
3. Sayers, E. W., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K. & Chetvernin, V. (2009). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 37, D5-15.
4. Horn, F., Vriend, G. & Cohen, F. E. (2001). Collecting and harvesting biological data: the GPCRDB and NucleaRDB information systems. Nucleic Acids Res 29, 346-9.
5. Dolk, E., van der Vaart, M., Lutje Hulsik, D., Vriend, G., de Haard, H., Spinelli, S., Cambillau, C., Frenken, L. & Verrips, T. (2005). Isolation of llama antibody fragments for prevention of dandruff by phage display in shampoo. Appl Environ Microbiol 71, 442-50.
6. Dolk, E., van Vliet, C., Perez, J. M., Vriend, G., Darbon, H., Ferrat, G., Cambillau, C., Frenken, L. G. & Verrips, T. (2005). Induced refolding of a temperature denatured llama heavy-chain antibody fragment by its antigen. Proteins 59, 555-64.
7. Folkertsma, S., van Noort, P., Van Durme, J., Joosten, H. J., Bettler, E., Fleuren, W., Oliveira, L., Horn, F., de Vlieg, J. & Vriend, G. (2004). A family-based approach reveals the function of residues in the nuclear receptor ligand-binding domain. J Mol Biol 341, 321-35.
8. Folkertsma, S., van Noort, P. I., de Heer, A., Carati, P., Brandt, R., Visser, A., Vriend, G. & de Vlieg, J. (2007). The use of in vitro peptide binding profiles and in silico ligand-receptor interaction profiles to describe ligand-induced conformations of the retinoid X receptor alpha ligand-binding domain. Mol Endocrinol 21, 30-48.
9. Kuipers, R. K. P., Joosten, H. J., Verwiel, E., Paans, S., Akerboom, J., van der Oost, J., Leferink, N. G., van Berkel, W. J., Vriend, G. & Schaap, P. J. (2009). Correlated mutation analyses on super-family alignments reveal functionally important residues. Proteins 76, 608-16.
10. Narayanan, B., Niu, W., Joosten, H. J., Li, Z., Kuipers, R. K., Schaap, P. J., Dunaway-Mariano, D. & Herzberg, O. (2009). Structure and function of 2,3-dimethylmalate lyase, a PEP mutase/isocitrate lyase superfamily member. J Mol Biol 386, 486-503.
11. Oliveira, L., Hulsen, T., Lutje Hulsik, D., Paiva, A. C. & Vriend, G. (2004). Heavier-than-air flying machines are impossible. FEBS Lett 564, 269-73.
12. Oliveira, L., Paiva, A. C. & Vriend, G. (2002). Correlated mutation analyses on very large sequence families. Chembiochem 3, 1010-7.
13. Oliveira, L., Paiva, P. B., Paiva, A. C. & Vriend, G. (2003). Identification of functionally conserved residues with the use of entropy-variability plots. Proteins 52, 544-52.
14. Oliveira, L., Paiva, P. B., Paiva, A. C. & Vriend, G. (2003). Sequence analysis reveals how G protein-coupled receptors transduce the signal to the G protein. Proteins 52, 553-60.
15. Skrabanek, L., Murcia, M., Bouvier, M., Devi, L., George, S. R., Lohse, M. J., Milligan, G., Neubig, R., Palczewski, K., Parmentier, M., Pin, J. P., Vriend, G., Javitch, J. A., Campagne, F. & Filizola, M. (2007). Requirements and ontology for a G protein-coupled receptor oligomerization knowledge base. BMC Bioinformatics 8, 177.

16. Van Durme, J., Horn, F., Costagliola, S., Vriend, G. & Vassart, G. (2006). GRIS: glycoprotein-hormone receptor information system. Mol Endocrinol 20, 2247-55.
17. Venselaar, H., Joosten, R. P., Vroling, B., Baakman, C. A., Hekkelman, M. L., Krieger, E. & al., e. (2009). Homology modelling and spectroscopy, a never-ending love story. Eur Biophys J.
18. Ye, K., Vriend, G. & AP, I. J. (2008). Tracing evolutionary pressure. Bioinformatics 24, 908-15.
19. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247, 536-40.
20. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389-402.
21. Vriend, G. & Sander, C. (1991). Detection of common three-dimensional substructures in proteins. Proteins 11, 52-8.
22. Rao, S. T. & Rossmann, M. G. (1973). Comparison of super-secondary structures in proteins. J Mol Biol 76, 241-56.
23. Holm, L., Ouzounis, C., Sander, C., Tuparev, G. & Vriend, G. (1992). A database of protein structure families with common folding motifs. Protein Sci 1, 1691-8.
24. Vriend, G. (1990). WHAT IF: a molecular modeling and drug design program. J Mol Graph 8, 52-6, 29.
25. Oliveira, L., Paiva, A. C. & Vriend, G. (1993). A model for G-protein coupled receptors. J Comp Aided Mol Des 7, 649-58.
26. Folkertsma, S., van Noort, P. I., Brandt, R. F., Bettler, E., Vriend, G. & de Vlieg, J. (2005). The nuclear receptor ligand-binding domain: a family-based structure analysis. Curr Med Chem 12, 1001-16.
27. Horn, F., Lau, A. L. & Cohen, F. E. (2004). Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors. Bioinformatics 20, 557-68.
28. Kuipers, R. K. P. & van de Bergh, T. (2009). Validator and Mutator: novel 3DM modules for prediction of pathogenic variants applied to Fabry's desease. Manusscript in preparation.
29. Olivier, M., Eeles, R., Hollstein, M., Khan, M. A., Harris, C. C. & Hainaut, P. (2002). The IARC TP53 database: new online mutation analysis and recommendations to users. Hum Mutat 19, 607-14.
30. Beroud, C. & Soussi, T. (2003). The UMD-p53 database: new mutations and analysis tools. Hum Mutat 21, 176-81.
31. Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5, 113.
32. Krissinel, E. & Henrick, K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr D Biol Crystallogr 60, 2256-68.
33. Alen, P., Claessens, F., Schoenmakers, E., Swinnen, J. V., Verhoeven, G., Rombauts, W. & Peeters, B. (1999). Interaction of the putative androgen receptor-specific coactivator ARA70/ELE1alpha with multiple steroid receptors and identification of an internally deleted ELE1beta isoform. Mol Endocrinol 13, 117-28.

34. Joosten, H. J., Han, Y., Niu, W., Vervoort, J., Dunaway-Mariano, D. & Schaap, P. J. (2008). Identification of fungal oxaloacetate hydrolyase within the isocitrate lyase/PEP mutase enzyme superfamily using a sequence marker-based method. Proteins 70, 157-66.
35. Teplyakov, A., Liu, S., Lu, Z., Howard, A., Dunaway-Mariano, D. & Herzberg, O. (2005). Crystal structure of the petal death protein from carnation flower. Biochemistry 44, 16377-84.
36. Dunwell, J. M., Purvis, A. & Khuri, S. (2004). Cupins: the most functionally diverse protein superfamily? Phytochemistry 65, 7-17.
37. Leferink, N. G., Heuts, D. P., Fraaije, M. W. & van Berkel, W. J. (2008). The growing VAO flavoprotein family. Arch Biochem Biophys 474, 292-301.
38. Leferink, N. G., Fraaije, M. W., Joosten, H. J., Schaap, P. J., Mattevi, A. & van Berkel, W. J. (2009). Identification of a gatekeeper residue that prevents dehydrogenases from acting as oxidases. J Biol Chem 284, 4392-7.
39. Baron, R., Riley, C., Chenprakhon, P., Thotsaporn, K., Winter, R. T., Alfieri, A., Forneris, F., van Berkel, W. J., Chaiyen, P., Fraaije, M. W., Mattevi, A. & McCammon, J. A. (2009). Multiple pathways guide oxygen diffusion into flavoenzyme active sites. Proc Natl Acad Sci U S A 106, 10603-8.
40. Mardis, E. R. (2006). Anticipating the 1,000 dollar genome. Genome Biol 7, 112.
41. Glover-Kerkvliet, J. (1994). p53 in 3-D. Environ Health Perspect 102, 1034-6.
42. Sedlacek, Z., Kodet, R., Poustka, A. & Goetz, P. (1998). A database of germline p53 mutations in cancer-prone families. Nucleic Acids Res 26, 214-5.
43. Stenson, P. D., Mort, M., Ball, E. V., Howells, K., Phillips, A. D., Thomas, N. S. & Cooper, D. N. (2009). The Human Gene Mutation Database: 2008 update. Genome Med 1, 13.

Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces.

Hanka Venselaar, Tim te Beek, Remko Kuipers,
Maarten Hekkelman, Gert Vriend

Abstract

Background

Many newly detected point mutations are located in protein-coding regions of the human genome. Knowledge of their effects on the protein's 3D structure provides insight into the protein's mechanism, can aid the design of further experiments, and eventually can lead to the development of new medicines and diagnostic tools.

Results

In this article we describe HOPE, a fully automatic program that analyzes the structural and functional effects of point mutations. HOPE collects information from a wide range of information sources including calculations on the 3D coordinates of the protein by using WHAT IF Web services, sequence annotations from the UniProt database, and predictions by DAS services. Homology models are built with YASARA. Data is stored in a database and used in a decision scheme to identify the effects of a mutation on the protein's 3D structure and function. HOPE builds a report with text, figures, and animations that is easy to use and understandable for (bio)medical researchers.

Conclusions

We tested HOPE by comparing its output to the results of manually performed projects. In all straightforward cases HOPE performed similar to a trained bioinformatician. The use of 3D structures helps optimize the results in terms of reliability and details. HOPE's results are easy to understand and are presented in a way that is attractive for researchers without an extensive bioinformatics background.

Background

The omics-revolution has led to a rapid increase in detected disease-related human mutations. A considerable fraction of these mutations is located in protein-coding regions of the genome and thus can affect the structure and function of that protein, thereby causing a phenotypic effect. Knowledge of these structural and functional effects can aid the design of further experiments and can eventually lead to the development of better disease diagnostics or even medicines to help cure patients. The analysis of mutations that cause the EEC syndrome, for example, revealed that some patients carry a mutation that disturbs dimerisation of the affected P63 protein [1]. This information has triggered a search for drugs (www.epistem.eu; [2]). In another case, the study of a mutation in the human hemochromatosis protein (HFE), which causes hereditary hemochromatosis, resulted in new insights that are now being used to develop novel diagnostic methods [3]. These and numerous other examples have highlighted the importance of using heterogeneous data, especially structure information, in the study of human disease-linked protein variants.

The data that can aid our understanding of the underlying mechanism of disease related mutations can range from the protein's three-dimensional (3D) structure to its role in biological pathways, or from information generated by mutagenesis experiments to predicted functional motifs. Collecting all available information related to the protein of interest can be challenging and time-consuming. It is a difficult task to extract exactly those pieces of information that can lead to a conclusion about the effects of a mutation. Several online Web servers exist that offer help to the (bio)medical researcher in predicting these effects. These servers use information from a wide range of sources to reach conclusions about the pathogenicity of a mutation. The PolyPhen server, for example, is widely used by researchers to predict the possible impact of an amino acid substitution on the structure and function of human proteins [4]. PolyPhen combines a subset of the UniProt sequence features, structural information (when available), and multiple sequence alignments in order to draw conclusions about the impact of a mutation [4]. SIFT, on the other hand, bases its mutation analysis purely on a multiple sequence alignment [5]. This server gives probability scores for each amino acid type at the position of interest to separate the harmless mutations from disease-causing ones. The ALAMUT software (http://www.interactive-biosoftware.com/) is widely used in human genetics research groups. It focuses on making many forms of software and databases available to their users. The ALAMUT system also automatically calls the PolyPhen Web server as part of its decision process. ALAMUT is not available as a Web server. PolyPhen, Sift and Alamut all have an excellent track-record make existing data accessible for (bio)medical scientist to aid them with the interpretation of mutational effects. We built on their strengths to produce the HOPE software that was written to optimally use the advantages of the novel tools of the e-Science era.

The recent increase in data types and data volumes has gone hand-in-hand with large efforts in bioinformatics that have led to numerous new databases and computational

methods, and in this era of e-Science, Web services provide on-demand access to these facilities [6-8]. The development of Web services facilitates the usage of external databases and methods in in-house developed software and eases software maintenance and development by out-sourcing logic to Web services. Web services have a series of advantages for the software developers:

- They save time by reusing program code;
- They tend to always be up-to-date;
- They are executed remotely, which gives access to large amounts of (free) CPU time, thus not overloading the local machine;
- No need to maintain in-house data and software collections.

Web services also have disadvantages:

- Source code of Web services often is not available;
- Web services are not guaranteed to always be available.

HOPE (Have (y)Our Protein Explained) is a next-generation web application for automatic mutant analysis. We have designed HOPE to explain the molecular origin of a disease related phenotype caused by mutations in human proteins. In this aspect HOPE resembles the aforementioned systems (PolyPhen, SIFT, ALAMUT). With HOPE we have taken the logical next step in the e-Science era in that the data gathering is done using Web services and DAS servers. Additionally, in HOPE we have taken a protein 3D structure centred approach. HOPE collects information from data sources such as the protein's 3D structure and the UniProt database of well-annotated protein sequences. For each protein this data is stored in a PostgreSQL-based information system. A decision scheme is used to process these data and to predict the effects of the mutation on the 3D structure and the function of the protein. A life-scientist friendly report is produced that explains and illustrates the effects of the mutation. This report is presented using an interface that is designed specifically for the intended user community of human genetics researchers. The report is enriched with figures that illustrate the effects of the mutation, while any residual bioinformatics jargon is linked to our in-house, online dictionary of bioinformatics jargon. The conclusions drawn in the report can be used to design follow-up experiments and eventually can lead to the development of better diagnostics or even medicines. Figure 1 illustrates the major steps of HOPE. We have tested HOPE on a series of mutations that we have previously analyzed manually. In all straightforward cases HOPE performed equally well as a trained protein structure bioinformatician.

Availability.

The HOPE Web server is freely available on http://www.cmbi.ru.nl/hope/.

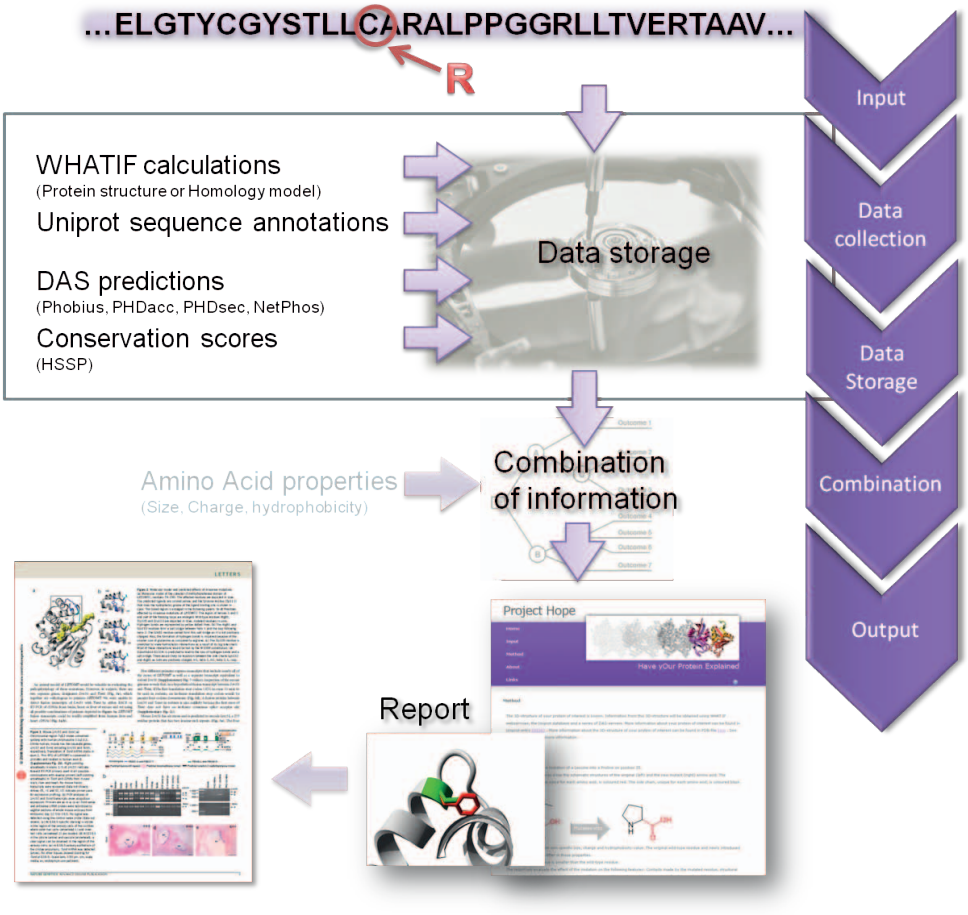


Figure 1. Overview of HOPE's process flow. The user submits a sequence and a mutation. HOPE will first collect information from a wide range of information sources. These sources include: WHAT IF for structural calculations on either the PDB file or a homology model that was build by YASARA, HSSP for conservation scores, DAS-servers for sequence-based predictions and Uniprot for sequence annotations. The data is stored in HOPE's information system. The data is combined with the known properties of the amino acids in a decision schedule. The result is a report shown on the HOPE website that will focus on the effect of the submitted mutation on the 3D-structure of the protein. The text and figures can be used in articles and publications.

Results and discussion

Input

The intended users of HOPE are life scientists who neither routinely use protein structures nor bioinformatics in their research. Therefore, both HOPE's input and its results are designed to be intuitive and simple, and all software used will run with default settings so

that the user neither needs to set parameters nor needs to read documentation. Actually, the user will not even know which software runs in the background. The interface of HOPE is a website that enables the user to submit a sequence and a mutation. The user can indicate the mutated residue and the new residue type by simple mouse-clicks. Figure 2 shows the input screen, filled with an example protein sequence and a mutation.

Figure 2. HOPE's input screen. The user can submit a sequence of interest and indicate the mutated residue with two simple mouse-clicks. In this example HOPE will analyze a leucine to aspartic acid mutation on position 25 of the plant protein Crambin.

Information retrieval

HOPE uses the submitted sequence as query for BLAST [9] searches against both the UniProt database [10] and the Protein Data Bank [11]. The search against the UniProt database identifies the protein's UniProt entry and the accession code of the protein, unique identifier that is used later in the process to obtain DAS-predictions. Alternatively, it is possible to submit this accession code directly. The BLAST search against the PDB is required to find the protein's structure or a possible template for homology modelling. HOPE uses the actual PDB-file when it contains the residue that is to be mutated and when it is 100% identical with the submitted sequence. HOPE identifies among multiple 100% hits the best structure for analysis based on resolution, experimental method, and length of the protein covered in the PDB (a full protein is preferred over a fragment). Nowadays, 20% of the human sequences available from SwissProt have a (partly) known structure and for another 30% a homology model can be build. To be able to build a homology model, the BLAST results should contain the equivalent location of the mutation and the percentage sequence identity should fall above the Sander and Schneider curve shown in Figure 3.

Homology modelling is performed using the Twinset version of YASARA which contains an automatic homology modelling script that requires only a sequence as input [12]. The script fully automatically performs the modelling process including sequence alignment, loop building, side chain modelling, and energy minimization. This script was the top contestant in the CASP8 modelling competition in terms of model detail accuracy [13].

The structure of the protein of interest, either a PDB-file or a homology model, is analyzed using WHAT IF Web services [7]. These services can calculate a wide range of structural features (e.g. accessibility, hydrogen bonds, salt bridges, ligand or ion interactions, mutability, variability, etc). When neither a 3D structure nor a possible modelling template is available, HOPE cannot use structural information and will instead base its conclusions only on the sequence related data, and published mutation and variation results.

The UniProt database (http://www.uniprot.org/) is used for the retrieval of features that can be mapped on the sequence [14]. This information includes the location of active sites, transmembrane domains, secondary structure, domains, motifs, experimental information, and sequence variants. The UniProt accession code is used to retrieve data from a series of DAS-servers for sequence based predictions such as possible phosphorylation sites. The DAS-servers form a widely used system for biological sequence annotation [15]. The conservation score of the mutated residue is calculated from a HSSP multiple sequence alignment [16].

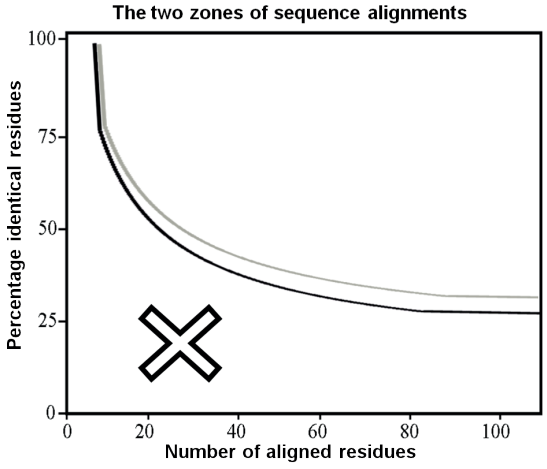


Figure 3. The two zones of sequence alignment identity that indicate the likelihood of adopting similar structures. Two aligned sequences are highly likely to have similar folds if their length and percentage sequence identity fall in the region above the threshold (black line). HOPE will build a homology model when the identity between the template and submitted sequence falls in this zone. In case the sequence identity is less than 5% above this threshold (grey line) HOPE will build a model but will also warn that the model is based on a template with low identity. The region below the threshold (indicated with a cross) indicates the zone where inference of structural similarity cannot be made, thus making it difficult to determine if model building will be possible. (Figure free after Sander and Schneider [25])

Data storage in HOPE

Information obtained from the protein structure or model, the UniProt record, and the DAS-predictions is stored in a protein-specific information system based on the PostgreSQL database system. One new information system is produced for each submitted protein. Differences in the protein sequence might exist between data sources, for example sequences from UniProt often contain the signal peptide while the sequences stored in the PDB tend to lack these residues. Therefore, sequences obtained from different sources are aligned using ClustalW. This enables us to transfer information to the residue of interest without the need to deal with the residue numbering problem that results from these sequence differences. Protein features are stored in the information system on a per-residue basis, and can have one of the following four data-types:

- Contacts: Interaction of the residue with another entity; for example DNA, a metal-ion, a ligand, hydrogen bond, disulfide bond, salt bridge;
- Variable features: Type with a value: for example, accessibility or torsion angle;
- Fixed features: Labels a residue (or stretch of residues) with a feature without a value. This indicates that the residue is located in a domain or motif (for example a residue can be part of the active site or in a transmembrane region);
- Variants: Mutations or other variations in sequence known at this position; for example splice variants, mutagenesis sites, SNPs.

After a user request has triggered the generation of an information system for the protein of interest, the system for this protein is kept on disk for one month just in case the same user (or another user for that matter) requests information about other mutations in the same molecule. After one month every system is thrown away to ensure that conclusions are never based on outdated information. So, there does not really exist a HOPE database as all HOPE's data is, in total agreement with e-Science paradigms, scattered over the internet, and is each time combined upon request.

Decision scheme

The decision scheme in HOPE uses all collected information combined with known properties of the wild-type and mutated amino acid, such as size, charge, and hydrophobicity, to predict the effect of the mutation on the protein's structure and function. The scheme consists of six parts that each correspond to a paragraph in the output. Each part analyzes the effect of the mutation on one of the following aspects of the residue:

- Contacts. Any interaction with other molecules or atoms, like DNA, ligans, metals, etc, but also hydrogen bonds, disulfide bridges, ionic interactions, etc;
- Structural domain. Any part of the protein with a specific name (and often function), such as domains, motifs, regions, transmembrane domains, repeats, zinc fingers, etc;

- Modifications. Features that do not directly influence the structure of the protein but might influence post-translational processes like phosphorylation.
- Variants. Known polymorphisms, mutagenesis sites, splice variants, etc;
- Conservation: The relative frequency of an amino acid type at each position taken from a multiple sequence alignment.
- Amino acid properties: The differences in the known properties of the wild-type and mutant residue (size, charge, hydrophobicity).

HOPE will produce its conclusions for each of these six aspects separately. For example, a residue can be located in a transmembrane domain and also be important for ligand interaction. HOPE will in this case produce a paragraph about the effect of the mutation on the contacts and a separate paragraph describing the effect of the mutation on the structural location, in this example the transmembrane domain.

Some types of information can be obtained from multiple sources, which are not equally reliable. Experimentally determined features and calculations performed on the 3D coordinates are more likely to be correct than any prediction. For example, transmembrane domains can be predicted by a DAS-server which normally will produce less reliable results than the annotations in UniProt. Therefore, HOPE ranks the information and uses the most accurate source available for its conclusions. WHAT IF calculations are preferred, followed by UniProt annotations, and DAS predictions are used only when neither WHAT IF nor UniProt data are available. In case no information about the mutated residue is found, HOPE will show a conclusion based only on biophysical characteristics between the wild type and mutant amino acid type. The conservation score is obtained either from the HSSP database that holds multiple sequence alignments for all proteins in the PDB, or through the HSSP Web services if a PDB file is not available [16].

Output

The report focuses on the effect of the mutation on the 3D-structure, and is aimed at a specific audience in the field of (bio)medical science. It shows the methods used and the sources of the combined information. This can either be an analysis of the real structure or homology model, or a prediction based on the sequence. The results of the mutation analyses are illustrated with figures of the amino acids and, if available, figures and animations of the mutation in the structure. The HOPE output is rather extensive and way too large to put in print, in figure 4 we just show a small part of one mutation report. A series of examples of HOPE output is available at the “about” section of the HOPE pages.

A HOPE result consists of one HTML page that contains all results. This makes it easy for users to print the results, or to make their own Web-page with HOPE results for long-term storage.

Test cases

HOPE was validated in a series of collaborations with scientists from different fields of life sciences. Experiences from these real-world examples where used to design and adjust the decision scheme. So far, most mutation studies involved non-sense and missense mutations. Descriptions of these projects can be found at the HOPE website. The resulting reports often contain a molecular explanation of the observed phenotype that can suggest further experiments. The majority of these projects included the building of a homology model as in most cases no 3D-structure of the protein of interest was available.

We also validated HOPE's conclusions by comparing them with the output of PolyPhen and SIFT. Even though it is very difficult to compare the results from PolyPhen, SIFT, and HOPE, we can still draw a few general conclusions, that will be elaborated on in the following paragraphs.

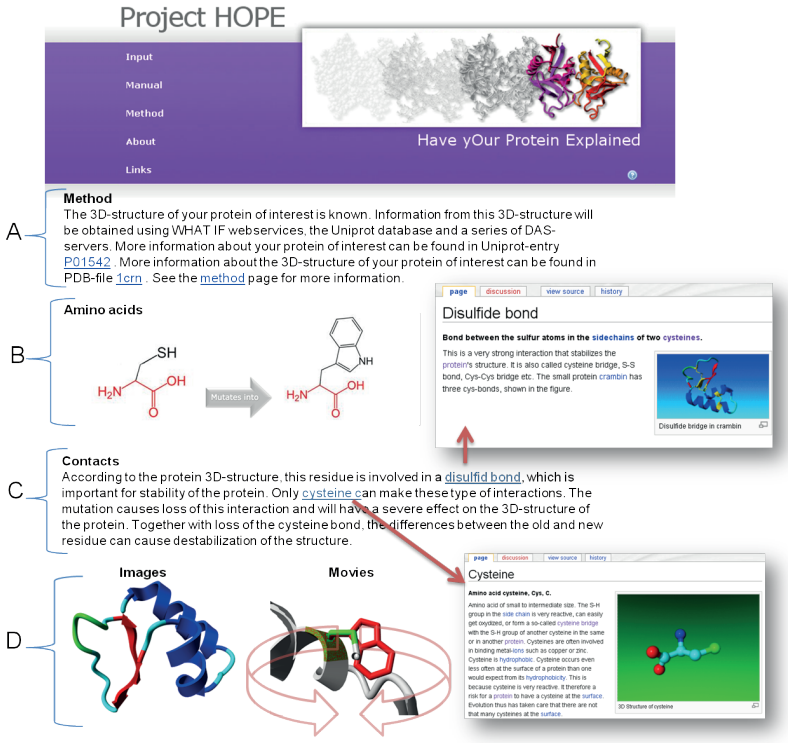


Figure 4. Example of HOPE's output. A simplified example of HOPE's output. A) Explanation of the used method (structure, modelling or predictions) and links to the relevant databases. B) Text and pictures that explain the differences between the wild-type and mutant residue. (Text is left out of this figure for clarity.) C) Paragraph of the report explaining the effect of the mutation on contacts made by the residue, a disulfid bond in this case. It contains a link to the wiki-entry "cysteine" and "disulfid bond". D) Images/animations that show the effect of the mutation on the structure.

Structure adds value

The use of a protein's 3D-structure or homology model increases the prediction quality in terms of reliability and detail. The possibility offered by the YASARA software to fully automatically build high quality homology models increases the number of sequences for which HOPE can use structure data. The protein structure, either a PDB-file or a homology model, can reveal information that currently cannot be predicted accurately from sequence alone, such as ionic interactions, ligand-contacts, etc.

The value of the extra information that HOPE can extract from a protein's structure or model is illustrated, for example, by the L320P and L347P mutations in ESRBB (see the "about" section of the HOPE website). All Web servers correctly predict the effect of these mutations as damaging for the protein. However, HOPE completes the story by an extensive explanation of the disturbing effect of prolines on alpha-helices. In cases for which no 3D structure data is available, the three Web servers seem to perform similarly albeit that Polyphen's output often tends to be scarce and a bit cryptic and SIFT's output is limited to conservation scores.

Biomedicist understandable results

HOPE's interface was designed especially for users that work in the (bio)medical sciences. Instead of displaying data in the form of detailed tables and numerical values, HOPE writes human readable reports that explain the structural and functional effects of the mutation, and illustrates this with figures and animations. When other Web servers list the effects of a mutation as "*Hydrophobicity change at buried site; normed accessibility: 0.00, hydrophobicity change: -2.7*". HOPE will instead report that "*the mutation introduces a less hydrophobic residue in the core of the protein which can destabilize the structure".* Many more examples of HOPE's readable output can be found at the "about" section of the HOPE website. HOPE's comprehensibility is improved by the Help-function that links difficult bioinformatics keywords to our own in-house dictionary based on Wikipedia's software. In this dictionary the user can find text, illustrations, and sometimes a short video-clip that explains the keyword.

Conclusions

Upon running 24 test cases, listed on the website, we realised that the present version of HOPE is useful and reliable in analysing point mutations. The next generation of HOPE will, however, need to reach a higher level of data integration to address more complicated cases. Some answer might be found only by combining the calculations with literature data and general knowledge of the protein's structure function relations. For example, PolyPhen predicts the N255D mutation in Kv1.1 (discussed in [17]) as being benign, while SIFT shows that this residue is 100% conserved. Combination of the conservation information with the fact that this residue is located in the voltage sensor of the channel can result in the hypothesis that the mutation disturbs the channel's voltage sensing mechanism. Such conclusions are still beyond the capabilities of today's Web servers, but the software design of HOPE will one day allow us to introduce the features needed to deal which these more complicated cases.

HOPE is an example of the new way of doing data- and software-intensive research in the era of eScience. Nowadays, the ongoing developments in experimental techniques like high-throughput sequencing will continue to produce large amounts of data and will therefore demand new, further automated approaches towards the analysis of these data. The eScience approach used will allow us to easily extend HOPE with more Web services, data sources, and DAS predictions when these become available. In the years to come HOPE can be extended with the possibility to analyze double-mutants, to quantitatively score the structural effects of the mutation and thereby provide the possibility to automatically rank candidate mutations that are the result of a sequence project, or to further improve the already user-friendly HOPE user interface.

Methods

The HOPE system is schematically shown in figure 5. The individual elements of his schema are described in the remainder of this section.

The HOPE website is implemented using the Wicket (http://wicket.apache.org/) web framework, which allows us to provide a fluent and responsive user experience. The web application is deployed on the GlassFish web application container (https://glassfish.dev.java.net/).

HOPE obtains information from different sources beyond our control. Therefore, the data gathering is set up as fail-safe as possible to handle service unavailability. Data is cached to speed up the process, reduce dependencies and to put less strain on external resources. The data-retention time is 30 days, after which time the data is renewed at the moment someone runs an analysis on the same sequence. The database scheme (available at the "about" section of the HOPE pages) is the result of an iterative design process using both Java and Hibernate to manage all data and to create the database tables. The database engine is PostgreSQL version 8.4.

The MRS BLAST version 4 Web service is used for most database searches with an e-value cut-off of 1e-5 and the low-complexity filter switched off [18]. This Web service (http://mrs.cmbi.ru.nl/mrsws/blast/wsdl) is backed by an in-house implementation of the standard BLAST algorithm. ClustalW version 2.0.10 is used for sequence alignments [19]. ClustalW is also offered as a Web service through MRS (http://mrs.cmbi.ru.nl/mrsws/clustal/wsdl).

WHAT IF Web services, accessible via http://wiws.cmbi.ru.nl/wsdl/, are used to calculate secondary structure (using DSSP [20]), accessibility values, structural fits of mutations, contacts with ligands or ions, salt bridges, disulfide bridges, and hydrogen bonds [21]. These calculations are performed either on the deposited PDB structure, or a homology model. Homology modelling is performed fully automatically using a locally installed WHAT IF & YASARA Twinset [12, 13]. This installation runs on a separate server, and is controlled through a Perl CGI script.

Sequence annotations are obtained from the UniProt database (http://www.uniprot.org/) [14] XML records. The obtained information includes sequence features such as active site, motifs, domains, variants and binding sites.

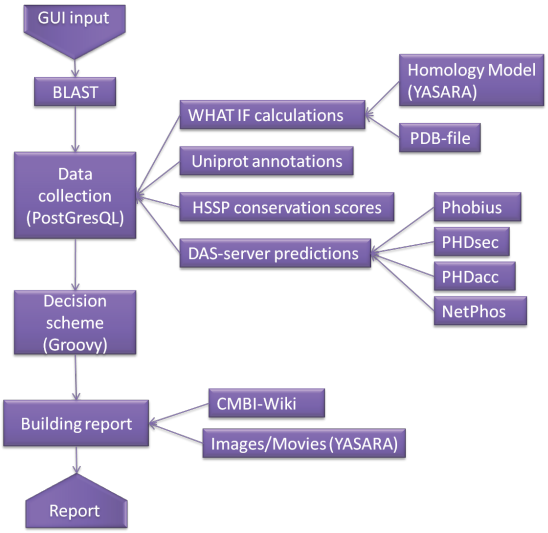


Figure 5. Detailed overview of HOPE's components. HOPE's input consists of the sequence and the mutation. The sequence is used for a BLAST search against the databases. Using the accession code (and PDB-file if available) HOPE can collect information from a series of information sources: WHAT IF calculations on the PDB-file or homology model built by YASARA, annotations in the Uniprot database, HSSP conservation scores and sequence-based predictions by DAS-servers. The information is combined in a decision scheme and a report is generated. This report is illustrated with pictures and animations and difficult keywords are linked to our own online dictionary.

Conservation scores are obtained from HSSP using the Web service for which the WSDL is available at http://mrs.cmbi.ru.nl/hsspsoap/wsdl. When a PDB deposited structure is available, the pre-calculated HSSP scores maintained at the CMBI are used. In case a homology model is available a DSSP file is generated for the homology model, which in turn is used to create a HSSP file. In case no structure or model is available, a HSSP file is generated using only the user sequence.

Distributed Annotation (DAS) servers [15, 22] are used to obtain predictions regarding transmembrane regions by Phobius [23], accessibilities by PHDacc [24], secondary structure by PHDsec [24], and phosphorylation sites by NetPhos [22].

The decision scheme is implemented in Groovy, a dynamic language that runs on the Java Virtual Machine (http://groovy.codehaus.org/). The simple Groovy language enables other users to design their own decision schemes and run a specific version of HOPE for their own purposes. The decision scheme is divided into separate branches targeted towards certain aspects of the mutant analysis, each producing a paragraph or sub-report. The decision scheme logic is separated from the phrases used to compose the report, for a cleaner separation in code and to allow for internationalization.

The HOPE report is presented on a self-contained webpage, allowing the user to save the page without breaking links and images. The user can bookmark the URL to perform the same mutant analysis at a later point in time, incorporating any newly available data. The output web pages are intended to be free from bioinformatics jargon. An online dictionary based on MediaWiki's software (http://www.mediawiki.org/) is used to explain bioinformatics-specific terms. JavaScript is used to link keywords on the webpage to articles present in the local MediaWiki instance. This functionality is available at any time via the omni-present blue help-button. Images and movies in the report are generated using the YASARA & WHAT IF Twinset.

Availability and Requirements

The full description of the design and implementation of the HOPE server is available from the "about" section of the HOPE pages. HOPE can be used freely and no licenses are required. The source code has been made open source and can be freely obtained from the HOPE website. HOPE uses Java, Groovy, and PostgreSQL; it has been implemented on a Linux system while care has been taken to avoid system dependencies.

Authors' contributions

HV was responsible for the biological implementation and validation of decision tree, and for the Wiki and the GUIs. TB designed and implemented the HOPE pipeline and most major elements; RK implemented the decision tree. MH and GV provided the WHAT IF Web services. HV and GV supervised the project. All authors read and approved the final manuscript.

Acknowledgements

The authors thank Elmar Krieger for his continuous support with the invaluable YASARA software. Barbara van Kampen en Wilmar Teunissen provided technical support. RK thanks NBIC for financial support. This work was part of the BioRange programme of the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI). GV acknowledges the EMBRACE project that is funded by the European Commission within its FP6 Programme, under the thematic area "Life sciences, genomics and biotechnology for health", contract number LHSG-CT-2004-512092.

References

1. Celli J, Duijf P, Hamel BC, Bamshad M, Kramer B, Smits AP, Newbury-Ecob R, Hennekam RC, Van Buggenhout G, van Haeringen A et al: Heterozygous germline mutations in the p53 homolog p63 are the cause of EEC syndrome. Cell 1999, 99(2):143-153.
2. Bykov VJ, Issaeva N, Shilov A, Hultcrantz M, Pugacheva E, Chumakov P, Bergman J, Wiman KG, Selivanova G: Restoration of the tumor suppressor function to mutant p53 by a low-molecular-weight compound. Nat Med 2002, 8(3):282-288.
3. Swinkels DW, Venselaar H, Wiegerinck ET, Bakker E, Joosten I, Jaspers CA, Vasmel WL, Breuning MH: A novel (Leu183Pro-)mutation in the HFE-gene co-inherited with the Cys282Tyr mutation in two unrelated Dutch hemochromatosis patients. Blood Cells Mol Dis 2008, 40(3):334-338.
4. Ramensky V, Bork P, Sunyaev S: Human non-synonymous SNPs: server and survey. Nucleic Acids Res 2002, 30(17):3894-3900.
5. Ng PC, Henikoff S: SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 2003, 31(13):3812-3814.
6. Pettifer S, Thorne D, McDermott P, Attwood T, Baran J, Bryne JC, Hupponen T, Mowbray D, Vriend G: An active registry for bioinformatics web services. Bioinformatics 2009, 25(16):2090-2091.
7. Hekkelman ML, Te Beek TA, Pettifer SR, Thorne D, Attwood TK, Vriend G: WIWS: a protein structure bioinformatics Web service collection. Nucleic Acids Res.
8. Bhagat J, Tanoh F, Nzuobontane E, Laurent T, Orlowski J, Roos M, Wolstencroft K, Aleksejevs S, Stevens R, Pettifer S et al: BioCatalogue: a universal catalogue of web services for the life sciences. Nucleic Acids Res, 38 Suppl:W689-694.
9. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. J Mol Biol 1990, 215(3):403-410.
10. The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res, 38(Database issue):D142-148.
11. Berman H, Henrick K, Nakamura H: Announcing the worldwide Protein Data Bank. Nat Struct Biol 2003, 10(12):980.
12. Krieger E, Koraimann G, Vriend G: Increasing the precision of comparative models with YASARA NOVA--a self-parameterizing force field. Proteins 2002, 47(3):393-402.
13. Krieger E, Joo K, Lee J, Raman S, Thompson J, Tyka M, Baker D, Karplus K: Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8. Proteins 2009, 77 Suppl 9:114-122.
14. Jain E, Bairoch A, Duvaud S, Phan I, Redaschi N, Suzek BE, Martin MJ, McGarvey P, Gasteiger E: Infrastructure for the life sciences: design and implementation of the UniProt website. BMC Bioinformatics 2009, 10:136.
15. Prlic A, Down TA, Kulesha E, Finn RD, Kahari A, Hubbard TJ: Integrating sequence and structural biology with DAS. BMC Bioinformatics 2007, 8:333.
16. Dodge C, Schneider R, Sander C: The HSSP database of protein structure-sequence alignments and family profiles. Nucleic Acids Res 1998, 26(1):313-315.
17. van der Wijst J, Glaudemans B, Venselaar H, Nair AV, Forst AL, Hoenderop JG, Bindels RJ: Functional analysis of the Kv1.1 N255D mutation associated with autosomal dominant hypomagnesemia. J Biol Chem, 285(1):171-178.
18. Hekkelman ML, Vriend G: MRS: a fast and compact retrieval system for biological data. Nucleic Acids Res 2005, 33(Web Server issue):W766-769.

19. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD: Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res 2003, 31(13):3497-3500.
20. Kabsch W, Sander C: Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983, 22(12):2577-2637.
21. Vriend G: WHAT IF: a molecular modeling and drug design program. J Mol Graph 1990, 8(1):52-56, 29.
22. Blom N, Gammeltoft S, Brunak S: Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol 1999, 294(5):1351-1362.
23. Kall L, Krogh A, Sonnhammer EL: A combined transmembrane topology and signal peptide prediction method. J Mol Biol 2004, 338(5):1027-1036.
24. Rost B: PHD: predicting one-dimensional protein structure by profile-based neural networks. Methods Enzymol 1996, 266:525-539.
25. Sander C, Schneider R: Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 1991, 9(1):56-68.

Correlated mutation analyses on super-family alignments reveal functionally important residues

Remko Kuipers, Henk-Jan Joosten, Eugene Verwiel, Sjoerd Paans,
Jasper Akerboom, John van der Oost, Nicole Leferink,
Willem van Berkel, Gert Vriend, Peter Schaap

Abstract

Correlated mutation analyses (CMA) on multiple sequence alignments are widely used for the prediction of the function of amino acids. The accuracy of CMA-based predictions is mainly determined by the number of sequences, by their evolutionary distances, and by the quality of the alignments. These criteria are best met in structure-based sequence alignments of large super-families. So far, CMA-techniques have mainly been employed to study receptor interactions. The present work shows how a novel CMA tool, called Comulator, can be used to determine networks of functionally related residues in enzymes. These analyses provide leads for protein engineering studies that are directed towards modification of enzyme specificity or activity. As proof of concept, Comulator has been applied to four enzyme super-families: the isocitrate lyase/phoshoenol-pyruvate mutase superfamily, the hexokinase super-family, the RmlC-like cupin super-family, and the FAD-linked oxidases super-family. In each of those cases networks of functionally related residue positions were discovered that upon mutation influenced enzyme specificity and/or activity as predicted. We conclude that CMA is a powerful tool for redesigning enzyme activity and selectivity.

Introduction

Proteins evolve within a framework of functional constraints that limit substitutions at individual positions in the sequence. The results of these constraints can be detected in large multiple sequence alignments as evolutionary fingerprints. Co-evolution of the amino acids at two distinct alignment positions, for example, is a result of functional constraints that force compensating mutations for specific residue changes. This co-evolution of residue positions can be detected by correlated mutation analyses (CMA) algorithms. Although the concept of correlated mutations is rather straightforward, their unambiguous detection proved more difficult. Therefore, several algorithms have been developed that are able to screen alignments for correlated mutations [1,2]. These methods are mostly used for the prediction of contacts between residues. Contact predictions can reveal inter-molecular protein-protein interactions [3-5], or intra-molecular interactions that in turn can be used for protein structure predictions [6]. In 1993 we introduced the idea that CMA is better suited for the detection of functionally related residues [7]. Later, using the GPCR protein super-family, we indeed showed that residue positions with common function tend to stay conserved, and when they do change they do so simultaneously [8]. Nevertheless, the number of articles that describe the utilization of CMA for detection of functionally related residues is still very limited. There are some examples where CMA was used to identify ligand-receptor interactions sites [8-10] and recently two papers appeared in which CMA was successfully used to detect residue positions important in multi-drug resistance of the HIV-1 protease [10,11].

Unambiguous detection of functionally related residues by CMA requires a reliable, large super-family alignment. We have recently designed the 3DM software (manuscript in preparation) that can be used to rapidly produce structure-based super-family multiple sequence alignments (MSAs). The Comulator software is a novel extension of this 3DM software suite, and was specifically designed for analysis of very large MSAs. Here it was used in protein engineering experiments to analyze the alignments of four super-families: (1) the isocitrate lyase/phoshoenol-pyruvate mutase (ICL/PEPM) super-family, in which we could selectively remove the specificity for one of its substrates, (2) the hexokinase (HK) super-family, in which we could properly predict the allowed subset of allowed residues in a saturation mutagenesis experiment, (3) the FAD-linked oxidases (FAD-O) super-family in which we could predict compensating mutations for loss-of-function mutants, and (4) the RmlC-like cupin (cupin) super-family (nomenclature according to the SCOP database [12]), in which we could improve the activity by designing a double mutant. The former two super-family MSAs were used to detect residues involved in substrate specificity, the latter two to predict compensating mutations for mutations that either decreased protein activity or protein stability. The Comulator CMA results agree in all four cases well with experiments; the latter two sets of experiments were produced by us and are here published for the first time.

The Comulator is available at http://www.3dmcsis.systemsbiology.nl/comulator/; this most likely is the first CMA software that is freely available in the internet. All underlying structure and sequence alignments, CMA results, mutations mined from literature, etc., are available from http://www.3dmcsis.systemsbiology.nl/.

Materials and methods

Super-family sequence alignment.

Protein structures belonging to 4 super-families were collected using the SCOP database combined with BLAST [13] searches in the PDB database. Super-family sequences were collected by searches in the NCBI database using the sequences of the super-family structures as query sequences. The super-family alignments were generated using 3DM. This software is described elsewhere (manuscript in preparation) and is only briefly described here. 3DM superposes the structures of proteins belonging to a super-family and so generates a structure-based sequence alignment that contains all sequence positions that are structurally conserved throughout the super-family. These so-called core positions are numbered sequentially, and these numbers (called 3DM-numbers) are used throughout this study. For each structure a profile is built by iteratively aligning sequences that are related to that structure. Finally, all sequences are aligned against the profile they are most similar to. In the final step the structure-based alignment is used for the generation of one large super-family alignment.

CM algorithm

The Comulator algorithm is derived from a method described for the detection of allosteric interactions in the nuclear receptor super-family [14]. The underlying method is known as the statistical coupling analysis method [15,16]. Equation 1 shows the similar method that was implemented in the Comulator.

$$CM(x,y) = \sum_{a=1}^{20} \sum_{b=1}^{20} |F(x,y,a,b)|\ N(x=a \wedge y=b)/400N$$

$$F(x,y,a,b) = \frac{N(y=b)}{N} - \frac{N(x=a \wedge y=b)}{N(x=a)}$$

In which x and y run over the residue positions in the MSA; $CM_{(x,y)}$ is the correlation score between the residue positions x and y; a and b run over the 20 amino acid types; N is the number of sequences in the MSA; $N_{(y=b)}$ is the frequency of residue type b at position y; $N_{(x=a)}$ is the frequency of residue type a at position x; and $N_{(x=a\text{^}y=b)}$ is the frequency of residue type b at position y in sequences where type a is observed at position x; $|F_{(x,y,a,b)}|$ is the absolute value of $F_{(x,y,a,b)}$. F will be negative if the amino acid b at position y is relatively more abundant in the

subset of sequences that has type a at position x ($N_{(x=a^y=b)}$/ $N_{(x=a)}$) than in the full alignment at position y ($N_{(y=b)}$/N). If the amino acid pair is observed less often in the subset than in the whole alignment the score F will be positive. Due to the summation of absolute values, pairs of residue types contribute more to the score $CM_{(x,y)}$ when their frequency deviates more from the average. In other words, a high F score is obtained when the residues at positions x and y tend to mutate in tandem. Obviously, a pair of fully conserved positions gets a score of zero. Comulator calculates CM scores for all possible alignment position pairs. The resulting scores are visualized in heat-maps that are incorporated in interactive HTML pages (fig. 1) in which all squares are hyperlinked to the underlying raw data, including the alignment.

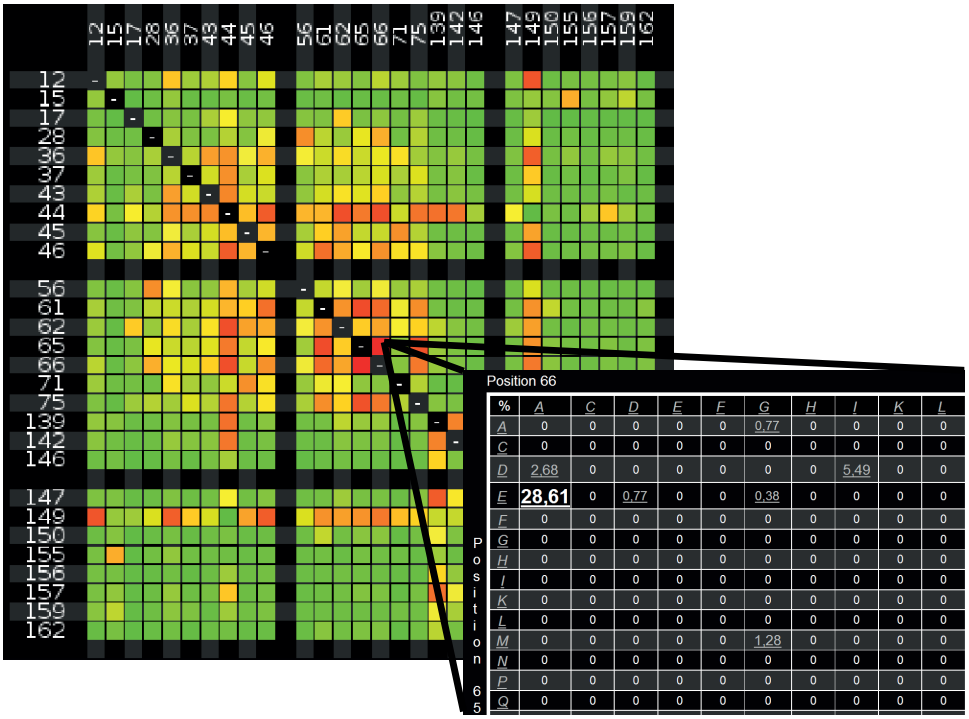


Figure 1. Heat-map of the ICL/PEPM super-family. Only residue positions are shown that have at least one correlation score above a cut-off (CMA score > 0.8). Squares are coloured from green (low correlation) to red (high correlation). The inset shows an example of hyperlinked information. In the insert the top-left corner of the residue pair frequency table is shown for the position pair (65,66).

Comulator website

The WWW based version of Comulator accepts as input aligned sequences in Fasta or ClustalW format. The input alignments are visualized similarly as for 3DM derived alignments including an alignment positions numbering scheme. The same numbering scheme is applied to the CM heat-map. If a sequence file contains a Swiss-Prot ID then the results are also automatically linked to the corresponding Swiss-Prot data file.

Mutation		*QuickChange primers*
3DM	*PfPGI*	
P27A	P132A	FW: (5'- GTAGTTTATGTTCCCGCCTATTGGGCTCATAGG -3') RV: (5'- CCTATGAGCCCAATAGGCGGGAACATAAACTAC -3')
Y28G	Y133G	FW: (5'- GTAGTTTATGTTCCCCCCGGTTGGGCTCATAGGACGG -3') RV: (5'- CCGTCCTATGAGCCCAACCGGGGGGAACATAAACTAC -3')
P27A/ Y28G	P132A/ Y133G	FW: (5'- GTAGTTTATGTTCCCGCCGGTTGGGCTCATAGGACGG -3') RV: (5'-CCGTCCTATGAGCCCAACCGGCGGGAACATAAACTAC -3')
P27E/ Y28G	P132E/ Y133G	FW: (5'- GTAGTTTATGTTCCCGAAGGTTGGGCTCATAGGACGG -3') RV: (5'- CCGTCCTATGAGCCCAACCTTCGGGAACATAAACTAC -5')
P27R/ Y28G	P132R/ Y133G	FW: (5'- GTAGTTTATGTTCCCCGCGGTTGGGCTCATAGGACGG -3') RV: (5'- CCGTCCTATGAGCCCAACCGCGGGGAACATAAACTAC -5')

Table 1. Primers used for the mutagenesis studies of pgiA. Both the 3DM alignment position numbering (1st column) and the number of the corresponding position in the ORF of phoshoglucose isomerase from P. furiosus (2nd column) are indicated.

Mutagenesis, over-expression and purification of phospho-glucose isomerase from Pyrococcus furiosus.

The cloning of *pgiA* is described by Verhees *et al*. [17]. Mutants were generated with the QuikChange Site-Directed Mutagenesis Kit (Stratagene, USA) following the manufacturer's instructions with the following adaptations: 25 PCR cycles were applied, and the PCR mixture was incubated with *Dpn*I for 4 to 8 hours at 37 °C. Mutants and primers used for mutagenesis are listed in table I. Mutants were verified by sequencing (Baseclear, Leiden, The Netherlands).

E. coli strain BL21(DE3) containing the tRNA accessory plasmid pRIL (Stratagene) carrying the concerning plasmid was routinely grown in 1 liter Luria Bertani medium (LB-medium) with kanamycin and chloramphenicol at 37 °C until an OD_{600} of 0.5 was obtained. Isopropyl-β-D-thiogalactopyranoside (IPTG) was added to a final concentration of 0.1 mM and the culture was further incubated for 8 hours under the same conditions. Cells were harvested by centrifugation (3,800 *g* at 4 °C for 20 min), resuspended in 10 ml lysis buffer (20 mM Tris·HCl, pH 8.0) and sonicated for 5 min at 4 °C. The cell extract was clarified by centrifugation (37,000 *g* at 4 °C for 20 min). *E. coli* proteins were denatured by incubating the cell extract at 70 °C for 30 min, and pelleted by centrifugation (37,000 *g* at 4 °C for 20 min). PGI was purified to homogeneity using FPLC: the supernatant was loaded onto a Q-Sepharose Fast Flow column (GE Healthcare, USA) pre-equilibrated with 20 mM Tris·HCl (pH 8.0). Proteins were eluted by a linear gradient of 0.0 to 1.0 M NaCl in 20 mM Tris·HCl (pH 8.0). Fractions containing PGI were pooled, concentrated, and loaded on a Superdex 200 GL column running in 20 mM Tris·HCl containing 125 mM NaCl.

Enzyme activity of the PGI mutants with fructose 6-phosphate was determined at 50°C as described previously[17] with the following adaptations: 20 mM Tris·HCl pH 7.0 was used, and the protein samples were pre-incubated with 50 mM EDTA at 50 °C for 20 minutes to ensure complete metal depletion [18]. Activity was measured with $MnCl_2$ in excess over EDTA to ensure enzyme saturation.

Mutation		*QuickChange primers*
3DM	*AtGALDH*	
36	L56H	FW: (5'-CCCGTTGGATCGGGTCACTCGCCTAATGGGATTG-3') RV: (5'-CAATCCCATTAGGCGAGTGACCCGATCCAACGGG-3')
78	A113G	FW: (5'-CTCTTCAGAACTTTGGCTCCATTAGAGAGCAG-3') RV: (5'-CTGCTCTCTAATGGAGCCAAAGTTCTGAAGAG-3')
91	V126G	FW: (5'-GGTGGTATTATTCAGGGTGGGGCACATGGGAC-3') RV: (5'-GTCCCATGTGCCCCACCCTGAATAATACCACC-3')

Table 2. Primers used for the mutagenesis Studies of AtGALDH. Both the 3DM alignment position numbering (1st column) and the number of the corresponding position in the ORF of L-galactono-1,4-lactone dehydrogenase from Arabidopsis thaliana (2nd column) are indicated.

Expression, purification, and mutagenesis of L-galactono-1,4-lactone dehydrogenase from Arabidopsis thaliana

The cDNA encoding mature L-galactono-1,4-lactone dehydrogenase (GALDH) from *A. thaliana* has been cloned previously to yield pET-GALDH-His_6 [REF]. The GALDH mutants used in this study were constructed using pET-GALDH-His_6 as template with the QuikChange method (Stratagene) using the primers listed in table 2. Successful mutagenesis was confirmed by automated sequencing.

For enzyme production *E. coli* BL21(DE3) cells, harbouring a pET-GALDH plasmid, were grown in 1 l LB-medium supplemented with 100 mg/ml ampicillin at 37°C until an OD_{600} of 0.7 was reached. Expression was induced by addition of 0.4 mM IPTG and the incubation was continued for 16 h at 37°C. The cells were harvested by centrifugation, resuspended in 5 ml lysis buffer (50 mM sodium phosphate, 300 mM NaCl, pH 7.4) and passed twice through a pre-cooled French Press (SLM Aminco) at 10 000 PSI. The resulting homogenate was centrifuged at 25 000 *g* for 30 min at 4°C to remove cell debris and the supernatant was loaded onto a HisGraviTrap column (GE Healthcare), equilibrated with 50 mM sodium phosphate, 300 mM NaCl, 45 mM imidazole, pH 7.4. Proteins were eluted with 50 mM sodium phosphate, 300 mM NaCl, 300 mM imidazole, pH 7.4 and saturated with FAD. Excess FAD and salt were removed by Biogel P-6DG size exclusion chromatography (BioRad) in 20 mM sodium phosphate, 0.1 mM DTT, pH 7.4. The amount of protein-bound FAD was determined from the ratio in absorbance at 280 and 450 nm (F-factor).

GALDH activity was routinely assayed by following the reduction of cytochrome *c* at 550 nm using a molar difference absorption coefficient (De_{550}) of 21 mM^{-1} cm^{-1} for reduced minus oxidized cytochrome *c* as described [19], with the modification that 1 mM FAD was included in the assay mixture. The thermal stability of GALDH was determined as reported earlier [19].

Results

Isocitrate lyase-like/Phosphoenolpyruvate mutase super-family.

The isocitrate lyase/PEP mutase super-family alignment contains enzymes that cover three of the six main enzyme families (EC codes) that all break a carbon-carbon in an oxalate-containing compound. All enzymes in this super-family share an α-β-barrel fold. The structure-based multiple sequence alignment contains 375 unique sequences. A network of nine residue positions with high CM scores was detected. These nine residue positions are found mainly but not exclusively surrounding the active site cleft. The detailed function of many of these residues is not yet known. However, these CM scores led to the discovery of a serine that is very specific for the oxaloacetate hydrolase (OAH) sub-family [20]. Mutating this serine to alanine, threonine, or proline (the most prevalent residues in other sequences in the alignment) did not significantly decrease the activity (k_{cat}), but had drastic effects on the affinity of OAH for its substrate oxaloacetate. This serine was used as a marker for family members with OAH activity, distinguishing OAH's from closely related paralogs that cannot convert oxaloacetate [20]. To show that this residue is indeed crucial for oxaloacetate recognition rather than for enzyme activity, we mutated this serine in the homologues petal death protein. This protein can convert a broad range of substrates including oxaloacetate. Indeed, when we mutated this serine we observed a 100 fold decrease in affinity for oxaloacetate, while the affinities for other substrates remained virtually unaffected [20]. Although OAH is very specific for oxaloacetate, it does have the potential to convert 2*R*,3*S*-2,3-dimethylmalate (DMM), albeit with poor efficiency. The S157A and S157P mutations in OAH actually improved the affinity of OAH for this substrate [20]. Recently we have isolated and characterized dimethyl-malate-lyase (DMML). DMML is closely related to OAH and has a proline instead of a serine at position 157. Mutating this proline to a serine shows the same behaviour with respect to k_{cat} and affinity [21].

Hexokinase super-family

The sequences of the hexokinase super-family can be divided in two main groups: The hexokinases that can phosphorylate a wide range of hexo-sugars, and glucokinases that specifically phosphorylate glucose. The Comulator detected a network of six highly correlated residue positions (3DM-numbers 46, 47, 52, 67, 82, and 88) that surround the active site (Fig. 2).

In humans/mammals glucokinases are hexokinases that often function in the liver. They are highly specific for glucose and work at high glucose concentrations. The 'other' hexokinases tend to have a broad range of substrates (e.g. alose, mannose, or glucosamine) and are observed in a wide variety of cell-types. In contrast to glucokinases, they function well at low substrate concentrations.

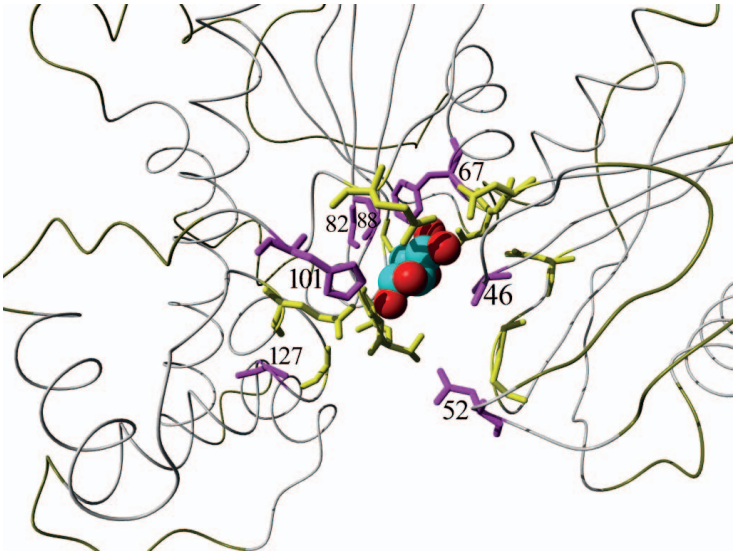


Figure 2. Tube presentation of a hexo kinase complexed with glucose (PDB accession code: 1BDG). Conserved residues (>90%) are shown in yellow and the network of highly correlating residues is shown in purple. The numbers assigned to the correlating positions are according to the 3DM numbering scheme.

The Comulator found a network of highly correlated positions that contact the substrate. 3DM-supported visual inspection of the 709 hexokinase sequences revealed that these six residue positions were conserved among the glucokinases. They were also conserved, but different among all 'other' hexokinases. The observed correlations almost perfectly separate the two main groups in the super-family. The fingerprints for these two groups are 46[A]47[C,G ,N]52[N]67[F]82[G]88[G] for the glucokinases, and 46[S]47[F,Y] 52[K]67[T]82[I]88[N] for the hexokinases. This clean separation suggests that these residue positions play a key role in determining the substrate specificities. Four of these positions (46, 47, 67, and 88) have been subject of a saturation mutagenesis experiment on the glucokinase of *E. coli* in a study by Miller [22]. *In vivo* selection in a glucokinase-deficient strain was used to find allowed substitutions at these positions. We analyzed all sequences in the MSA that have at least four of these six fingerprint residues in agreement with the glucokinase consensus (A,(CGN),N,F,G,G). Most of these sequences possess consensus residues at all six positions, but about ten percent differed at one or two positions. Despite the relatively low percentage of mutations, the results are still statistically meaningful because of the massive number

of sequences used in this study. Table 3 lists the residue types that were observed at these six fingerprint positions. The saturation mutagenesis experiments that were performed for four of these six residues are shown too. In 65 of 76 mutants we see that residues that are observed in the MSA also are detected in the saturation mutagenesis experiment and *vice versa*; in other words: residues not observed in the MSA were not observed experimentally.

	Ali	*Screen*	*Ali*	*Screen*	*Ali*	*Screen*	*Ali*	*Screen*	*Ali*	*Ali*
Nr.	*46*		*47*		*67*		*88*		*52*	*82*
A	WT	WT	-	-	-	-	-	-	+	+
C	-	+	+	+	-	+	-	-	-	-
D	-	-	-	+	-	-	-	-	-	-
E	-	-	-	-	-	-	-	-	-	-
F	-	-	+	+	WT	WT	-	-	-	-
G	+	+	WT	WT	-	-	WT	WT	-	WT
H	-	-	+	+	-	-	-	-	-	-
I	-	-	+	+	+	+	-	-	-	-
K	-	-	-	+	-	+	-	-	+	-
L	-	-	+	+	+	+	-	-	-	-
M	-	-	+	+	+	+	-	-	-	-
N	-	-	+	-	-	+	-	-	WT	-
P	+	+	-	-	+	+	-	-	-	-
Q	-	-	-	-	-	+	-	-	-	-
R	-	-	-	+	-	-	-	-	-	-
S	+	+	+	+	+	+	-	-	-	-
T	-	-	+	+	+	+	-	-	-	-
V	-	-	+	+	+	+	-	-	-	-
W	-	-	-	-	-	+	-	-	-	-
Y	-	-	-	+	-	-	-	-	-	-

Table 3. Residues observed at positions in the MSA that have at least four out of six glucokinase fingerprint residues according to the consensus, together with the results of saturation mutagenesis at four of these positions22. The residue numbers shown are 3DM-numbers. The corresponding numbers in the E. coli glucokinase are 64,65,101,140, 76,134, respectively. The 'Ali' columns show a plus sign if the amino acid type was detected at least once in the MSA at that position. Plus signs in the 'Screen' columns indicate that the saturation mutagenesis experiment showed that this residue type at that position produced viable protein. Minus signs indicate non-observed residue types. The colours green and red indicate whether CMA and experiment agreed or disagreed, respectively. WT indicates that that residue either is the consensus residue at that position, or is observed in the E coli wild type sequence. Not counting the WT cases, we observe agreement in 65 out of 76 cases.

FAD-binding domain of the vanillyl alcohol oxidase super-family

The vanillyl alcohol oxidase (VAO) flavoprotein family (FAD-O in SCOP database) is a large group of enzymes that catalyze a wide variety of oxidation-reduction reactions [23,24]. Members of this family share a characteristic domain topology, with a conserved N-terminal FAD-binding domain and a less well conserved C-terminal cap domain that determines the substrate specificity (fig. 3). Most members of the VAO flavoprotein family contain a covalently bound FAD cofactor. L-Galactono-1,4-lactone dehydrogenase (GALDH; SwissProt ac=Q8GY16) is a VAO-family member that is involved in the vitamin C biosynthesis pathway in plants. Structural information is neither available for GALDH nor for any close homolog. Consequently little is known about GALDH's active site or about the molecular basis for the non-covalent binding of the FAD cofactor, and a series of mutations was therefore made to obtain such information. Position 56 (36 in the 3DM alignment) is located in the so-called PP-loop that interacts with the pyrophosphate moiety of the FAD molecule [23]. In most family members with a covalently bound FAD, a histidine is observed at this position. Replacing the histidine at position 36 in covalent VAO-family members yielded either active proteins with non-covalently bound FAD, or inactive apo-proteins [25,26].

Figure 3. Crystal structure of 6-hydroxy-D-nicotine oxidase (PDB accession code: 2BVF). The FAD-binding domain is in green, the cap-domain is in red, the 8a-N1-histidyl-FAD cofactor is blue, and the residues with high CM scores are shown as grey ball models.

A 3DM alignment was constructed using 1152 sequences of (putative) VAO-family members. 3D protein structures are sparsely spread over this wide enzyme super-family. The 3DM alignment of the VAO-flavoprotein super-family comprises only the N-terminal FAD-binding domain due to a lack of structural conservation in the C-terminal cap domain. Alignment position 36 (a histidine in the VAO members with a covalently bound FAD) correlates well with positions 78, 91, and 92. Positions 78, 91 and 92 all are located in

the direct vicinity of the pyrophosphate moiety of the isoalloxazine ring of the flavin, with residue 78 being at hydrogen bonding distance of the reactive N5 locus (fig. 3).

Variant	k_{cat} *(s^{-1})*	K_m *(mM)*	*FAD binding (F-factor)*
Wild-type[a]	134 ± 5	0.17 ± 0.01	++ (8.0)
L36H[a]	32 ± 1	0.12 ± 0.01	++ (7.9)
A78G	116 ± 5	0.45 ± 0.03	++ (8.3)
V91G	62 ±	0.27 ±	± (14.1)
L36H/A78G	7.6 ± 0.2	0.15 ± 0.02	++ (8.2)
L36H/V91G	21 ±	0.15 ±	± (14.6)
A78G/V91G	49 ± 2	0.31 ± 0.06	+ (10.5)
L36H/A78G/V91G	<0.1	ND	-- (ND)

Table 4. Catalytic and FAD-binding properties of GALDH variants. Values for the wildt-type and L36H adapted from Leferink *et al.* [24]. ND: Not determined.

Among the VAO-family members with a covalently bound FAD, a histidine is favoured at position 36, and glycines are favoured at positions 78, 91, and 92. MurB reductases, which have a non-covalently bound FAD, favour a serine at position 36, and a leucine or alanine at position 78 and a methionine at position 91. GALDH contains a leucine at position 36 (Leu56), an alanine at position 78 (Ala113), and a valine at position 91 (Val126). Throughout the super-family, glycine is the preferred residue type at position 92 both for variants with a His at position 36 and for variants with a Leu at this position. Mutations studies were therefore started with the single mutants L36H, A78G, and V91G. These are three mutations that move the GALDH sequence in the direction of the consensus of family members with a covalently bound FAD. Covalently bound FAD was not observed in any of the GALDH variants.

Table 4 shows that the three single mutants have similar or worse K_m and k_{cat} values than the wild-type enzyme. Table 4 also shows that V91G poorly binds FAD. The Comulator residue type frequencies (table 5) indicate that all three single mutations led to situations with unfavourable amino acid combinations of the residue pairs shown to be important by the CMA.

With a His at position 36 the most abundant residue at position 78 is a Gly. When this Gly is added in the L36H/A78G double mutant, the catalytic rate is not improved, but the deteriorating effect on K_m of the A78G mutant is compensated (table 4). Note that the loss in k_{cat} in the L36H mutant is much stronger in the A78G background than in the

single mutant. With a His at position 36 the most abundant residue at position 91 is a Gly (table 5). When this Gly is added in the L36H/V91G double mutant, we see a similar effect as observed for the L36H/A78G double mutant, an improvement in the K_m but a loss of k_{cat}. With a Gly at position 91, the most abundant residue at position 78 is a Gly. When Gly78 was added in the A78G/V91G double mutant, a variant was obtained with catalytic properties comparable to both single mutants, but with much better FAD binding properties than V91G (table 4). Introducing a His at position 36 in the L36H/A78G/V91G triple mutant resulted in a mutant protein that is expressed as insoluble apo-protein that could not be purified in enough quantities to perform biochemical studies.

	36	*78*	*91*	*36/78*	*36/91*	*78/91*
WT	L	A	V	L/A (0.26)	L/V (0.09)	A/V (0.09)
L36H	H	A	V	H/A (0.35)	H/V (0.69)	A/V (0.09)
A78G	L	G	V	L/G (0.17)	L/V (0.09)	G/V (0.26)
V91G	L	A	G	L/A (0.26)	L/G (0.0)	A/G (0.61)
L36H/A78G	H	G	V	H/G (19.9)	H/V (0.69)	G/V (0.26)
L36H/V91G	H	A	G	H/A (0.35)	H/G (15.9)	A/G (0.61)
A78G/V91G	L	G	G	L/G (0.17)	L/G (0.0)	G/G (15.4)

Table 5. Comulator residue type frequencies of GALDH variants

RmlC-like cupin super-family

The alignment of the cupin super-family is the largest of the four superfamilies studied and contains 2097 sequences. The RmlC-like cupin super-family consists of proteins possessing a common β-barrel structure also known as a jelly roll fold. Although the proteins in this super-family are functionally diverse [27], most are enzymes of which the active site is located within the β-barrel. This active site often contains two histidines (3DM-numbers 10 and 31) that are conserved in approximately 80% of all sequences (fig. 4: yellow residues). Comulator revealed a network of highly correlating positions consisting of the alignment positions with 3DM-numbers 12, 19, 20, 26, 27,28, and 40 (fig. 4 magenta residues). The highest pair-wise CM score was detected for position pair 27 and 28. The alignment positions 26-28 form a structurally conserved surface loop in most members of the super-family.

The 3DM software automatically extracted many hundreds of mutations for the cupin super-family from the literature and stored them in the database to allow for rapid inspection. Intriguingly, the network of correlated positions has barely been mutated by anybody yet. Mutagenesis of position 28 in flavonol synthase from *Citrus unshiu* (G261A) resulted in 95% reduction of enzyme activity. Introduction of a proline at this same position 28 resulted in a completely inactive enzyme [29]. The residue at position 28 is located far away from the active site histidines (fig. 4) but nevertheless has been shown important for

activity. Because the residue positions 27 and 28 show the highest CMA value, we decided to study this pair of residues experimentally.

One of the best characterized members of the cupin super-family is the *Pyrococcus furiosus* PGI (PfPGI) [18,28,30]. Several crystal structures of this protein have been elucidated [18,28,30,31] and the reaction mechanism has been analyzed by mutagenesis, NMR, and EPR studies [30]. PfPGI has a tyrosine at position 28 whereas glycine is the most prevalent amino acid at this position in the cupin super-family (42%, see Fig. 5). The Y28G mutant (Y133G in PfPGI numbering) results in a 2.6 fold reduction of the activity (fig. 6). The most abundant residue at position 27 (that is highly correlated with 28) is glycine (present in 26% of all sequences); in PfPGI it is a proline residue. Analysis of the

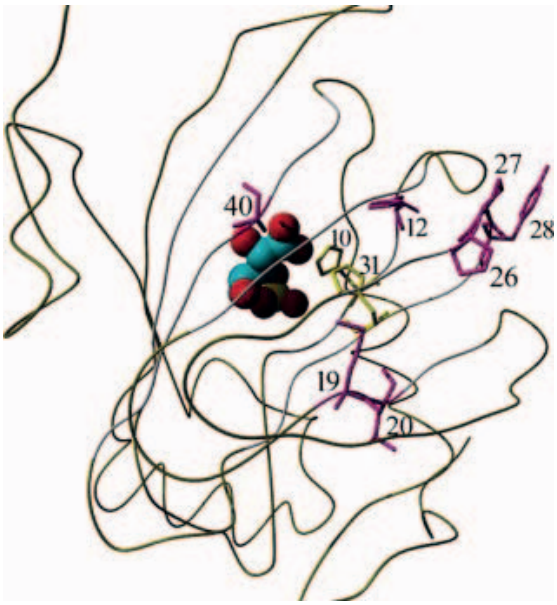


Figure 4. Tube representation of the 3D-structure of PfPGI from P. furiosus (PDB accession code: 1X82). The two conserved histidines are shown in yellow and the core positions with high CM scores are in magenta. The inhibitor 5-phospho-D-arabinonate represented as balls.

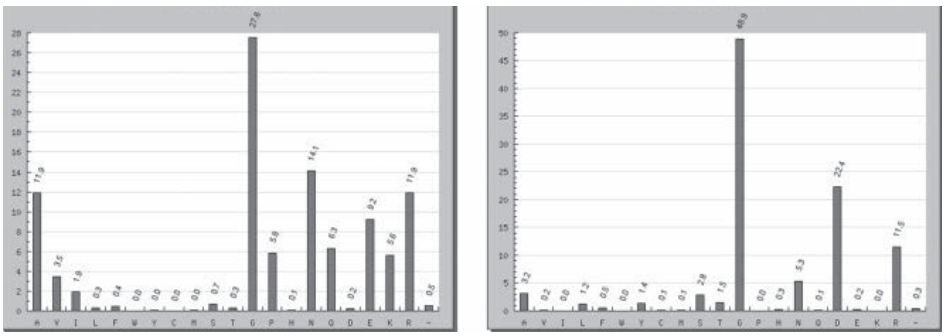


Figure 5. Bar graphs representing the amino acid distributions of positions 27 (left) and 28 (right). The x-axes lists the 20 different amino acids and the y-axes their percentages in the MSA.

Comulator heat-map reveals that the combination P27-Y28 occurs rarely (1.1%), whereas P27-G28 occurs in 4.4% of all sequences. However, when a glycine is observed at position 28, an alanine is by far the most prevalent residue at position 27 (23%). The double mutation P27A/Y28G not only regained the activity lost by the Y28G mutant, but even became twice as active as the wild-type enzyme. Obviously, if we had started with the P27A mutant and compensated it with Y28G, we would have obtained the same result (see fig 6).

Of the sequences with a glycine at position 28, 18.5% has a glutamate at position 27 and 18.2% an arginine. Both P27E and P27R can compensate for the loss of activity of Y28G regaining near WT activity (fig. 6).

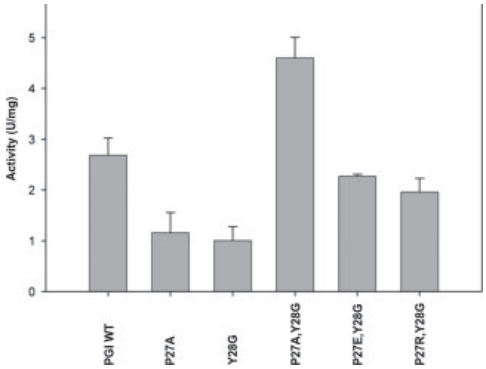


Figure. 6. Bar graphs representing activity of single or double mutants of wild type PfPGI. Numbering according to the 3DM numbering. The numbers in the amino acid sequence of PfPGI are 132 and 133, respectively.

Conclusions

Most enzyme engineering successes of the past decade have been accomplished via random mutagenesis, euphemistically called evolutionary approaches, while rational mutagenesis in terms of predicting one mutation at one position to achieve one phenotypic effect have lost terrain. The recent explosion in a series of high-throughput technologies, including sequencing, is enabling an even larger speed in the technical execution of these evolutionary approaches. It has often been observed, and this study adds one more observation, that optimal phenotypic effects tend to require a series of mutations to be introduced simultaneously, and many evolutionary approaches are optimized to achieve just that goal. Still, parallel random mutagenesis is technically limited to a handful of amino acid positions in the protein. It is therefore of paramount importance to select the positions well where these random mutations are going to be introduced. With CMA we can find groups of residues that are involved in the same function. And we have shown that this enables us to find combinations of mutations that improve catalysis rate or modify substrate specificity. It seems therefore that a major step forward can be made in enzyme engineering if the amino acid positions selected for combined randomisation are carefully selected from a CMA screen.

Our results also show that the combinatorial freedom at the positions detected by the CMA is limited so that full randomisation is not needed to harvest the complete combinatorial potential. Our mutation studies have shown that the combination of residues that can bring the desired phenotypic change in the enzyme often has already been tried in a different context, i.e. in another protein, so that the limited number of sequence finger prints obtained by CMA will be a good start for limited randomisation. Looking at the ease of today's gene synthesis approaches, we can imagine that this might be a new path towards semi rational enzyme engineering.

References

1. Halperin I, Wolfson H, Nussinov R. Correlated mutations: advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families. Proteins 2006;63:832-845.
2. Fodor AA, Aldrich RW. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. Proteins 2004;56:211-221.
3. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. Correlated mutations contain information about protein-protein interaction. J Mol Biol 1997;271:511-523.
4. Mintseris J, Weng Z. Structure, function, and evolution of transient and obligate protein-protein interactions. Proc Natl Acad Sci U S A 2005;102:10930-10935.
5. Kundrotas PJ, Alexov EG. Predicting residue contacts using pragmatic correlated mutations method: reducing the false positives. BMC Bioinformatics 2006;7:503.
6. Burkhard Rost & Sean O'Donoghue Sisyphus and protein structure prediction. Bioinformatics 1997;13:345-356.
7. L. Oliveira, A. Paiva, G. Vriend, J. A model for G-protein coupled receptors. Comp Aided Mol Des 1993;7:649-658.
8. Singer MS, Oliveira L, Vriend G, Shepherd GM. Potential ligand-binding residues in rat olfactory receptors identified by correlated mutation analysis. Receptors Channels 1995;3:89-95.
9. LinksPulim V, Bienkowska J, Berger B. LTHREADER: prediction of extracellular ligand-receptor interactions in cytokines using localized threading. Protein Sci 2008;17:279-292.
10. Garriga C, Pérez-Elías MJ, Delgado R, Ruiz L, Nájera R, Pumarola T, Alonso-Socas Mdel M, García-Bujalance S, Menéndez-Arias L. Mutational patterns and correlated amino acid substitutions in the HIV-1 protease after virological failure to nelfinavir- and lopinavir/ritonavir-based treatments. J Med Virol 2007;79:1617-1628.
11. Liu Y, Eyal E, Bahar I. Analysis of correlated mutations in HIV-1 protease using spectral clustering. Bioinformatics 2008;24:1243-1250.
12. Hubbard TJ, Ailey B, Brenner SE, Murzin AG, Chothia C. SCOP: a Structural Classification of Proteins database. Nucleic Acids Res 1999:27;254-256.
13. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
14. Shulman AI, Larson C, Mangelsdorf DJ, Ranganathan R. Structural determinants of allosteric ligand activation in RXR heterodimers. Cell 2004;116:417–429.
15. Lockless SW, Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. Science 1999:286;295-259.
16. Fodor AA, Aldrich RW. On evolutionary conservation of thermodynamic coupling in proteins. J Biol Chem 2004:279;19046-19050.
17. Verhees CH, Huynen MA, Ward DE, Schiltz E, de Vos WM, van der Oost J. The phosphoglucose isomerase from the hyperthermophilic archaeon *Pyrococcus furiosus* is a unique glycolytic enzyme that belongs to the cupin super-family. J Biol Chem 2001;276:40926-40932.
18. Berrisford JM, Akerboom J, Turnbull AP, de Geus D, Sedelnikova SE, Staton I, McLeod CW, Verhees CH, van der Oost J, Rice DW, Baker PJ. Crystal structure of *Pyrococcus furiosus* phosphoglucose isomerase. Implications for substrate binding and catalysis. J Biol Chem 2003;278:33290-33297.

19. Leferink NGH, Van den Berg WAM, Van Berkel WJH. L-Galactono-γ-lactone dehydrogenase from Arabidopsis thaliana, a flavoprotein involved in vitamin C biosynthesis. FEBS J 2008;275:713-726.
20. Joosten HJ, Han Y, Niu W, Du J, Vervoort J, Dunaway-Mariano D, Schaap PJ. Identification of Fungal Oxaloacetate Hydrolyase Within the Isocitrate Lyase/PEP Mutase Enzyme Superfamily Using a Sequence Marker Based Method. Proteins 2008;70:157-166
21. Narayanan B, Niu W, Joosten HJ, Kuipers RKP, Li Z, Schaap PJ, Dunaway-Mariano D, Herzberg O. Structure and Function of 2,3-Dimethylmalate Lyase, a PEP Mutase/Isocitrate Lyase Superfamily Member. Submitted.
22. Miller BG. The mutability of enzyme active-site shape determinants. Protein Sci 2007;16:1965-1968.
23. Fraaije MW, Van Berkel WJ, Benen JA, Visser J, Mattevi A. A novel oxidoreductase family sharing a conserved FAD-binding domain. Trends Biochem Sci 1998;23:206-207.
24. Leferink NGH, et al. The growing VAO flavoprotein family. Archives of Biochemistry and Biophysics 2008. In Press.
25. Caldinelli L, Iametti S, Barbiroli A, Fessas D, Bonomi F, Piubelli L, Molla G, Pollegioni L. Relevance of the flavin binding to the stability and folding of engineered cholesterol oxidase containing noncovalently bound FAD. Protein Sci. 2008;17:409-419.
26. Heuts DP, van Hellemond EW, Janssen DB, Fraaije MW. Discovery, characterization and kinetic analysis of an alditol oxidase from Streptomyces coelicolor. J Biol Chem 2007:282:20283-20291.
27. Dunwell JM, Purvis A, Khuri S. Cupins: the most functionally diverse protein super-family. Phytochemistry 2004:65;7-17
28. Berrisford JM, Akerboom J, Brouns S, Sedelnikova SE, Turnbull AP, van der Oost J, Salmon L, Hardré R, Murray IA, Blackburn GM, Rice DW, Baker PJ. The structures of inhibitor complexes of Pyrococcus furiosus phosphoglucose isomerase provide insights into substrate binding and catalysis. J Mol Biol. 2004;343:649-657.
29. Wellmann F, Lukacin R, Moriguchi T, Britsch L, Schiltz E, Matern U. Functional expression and mutational analysis of flavonol synthase from Citrus unshiu. Eur J Biochem. 2002:269;4134-4142.
30. Berrisford JM, Hounslow AM, Akerboom J, Hagen WR, Brouns SJ, van der Oost J, Murray IA, Michael Blackburn G, Waltho JP, Rice DW, Baker PJ. Evidence supporting a cis-enediol-based mechanism for Pyrococcus furiosus phosphoglucose isomerase.J Mol Biol 2006;358:1353-1366.
31. Hansen T, Oehlmann M, Schonheit P. Novel type of glucose-6-phosphate isomerase in the hyperthermophilic archaeon *Pyrococcus furiosus*. J Bacteriol 2001;183:3428-3435.

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

Remko Kuipers, Tom van den Bergh,
Henk-Jan Joosten, Ronald Lekanne dit Deprez,
Marcel Mannens, Peter Schaap

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.

Introduction

Due to the ease of today's gene sequencing methods, the relation between genes and corresponding diseases has been unraveled 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 predetermined 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.

Materials & Methods

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.

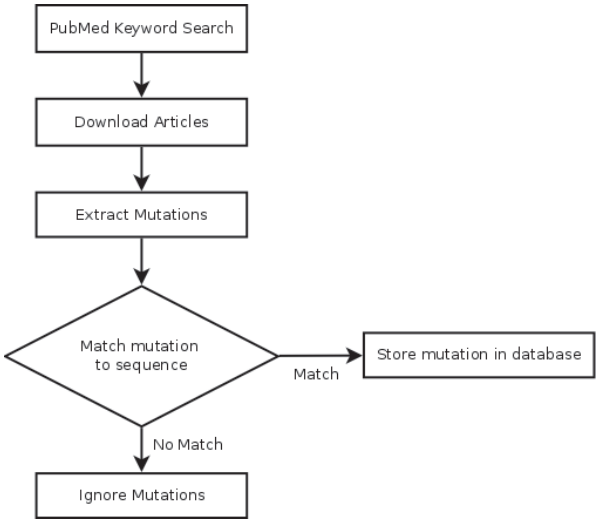


Figure 1. Schematic flowchart of Mutator program. See supplementary workflow S1 for the algorithm of Mutator and supplementary figure S2 for a more detailed flowchart of the algorithm.

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.

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. 3 & 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].

Results & Discussion

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-acetylgalactosaminidase (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.

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 shows that 80% of the reported mutations are at structurally conserved positions (core). Two regions outside the core are highly populated with Fabry related

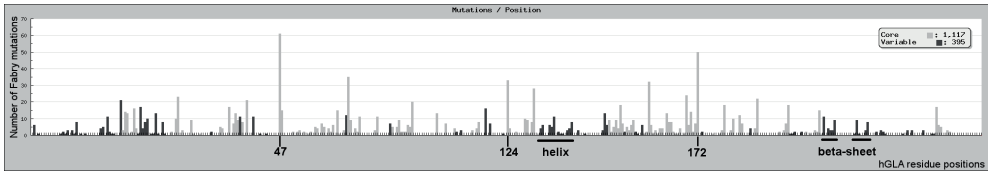


Figure 2. Number of independently reported Fabry disease related mutations per GLA residue position. Independently reported mutations detected at structural conserved positions (core) are in light grey. Structural non-conserved positions are in dark grey. More then 40 independently reported mutations 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 out of a total of 1512) are observed at those positions.

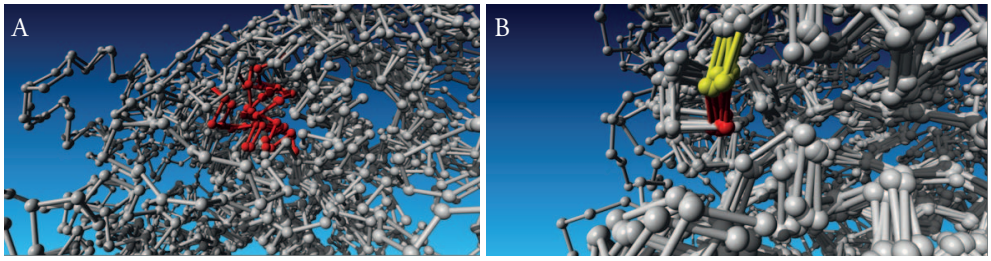


Figure 3. YASARA ball and stick backbone representation of seven superimposed protein structures of different sub-families of the alpha-amylase super-family. 3A: In red equivalent helices from the seven proteins. This helix is present in almost all super-family members, but could not be included in the 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.

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.

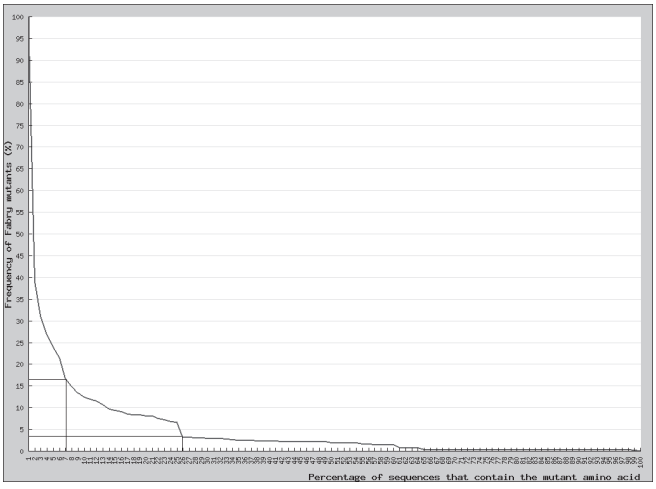


Figure 4. Correlation between the relative amino acid conservation (x-axis) and frequency of reported Fabry related mutations (y-axis). The x-axis represents the percentage of sequences that contains the mutated residue. The y-axis represents the percentage of the total number of fabry related mutation 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.

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 suggest 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.

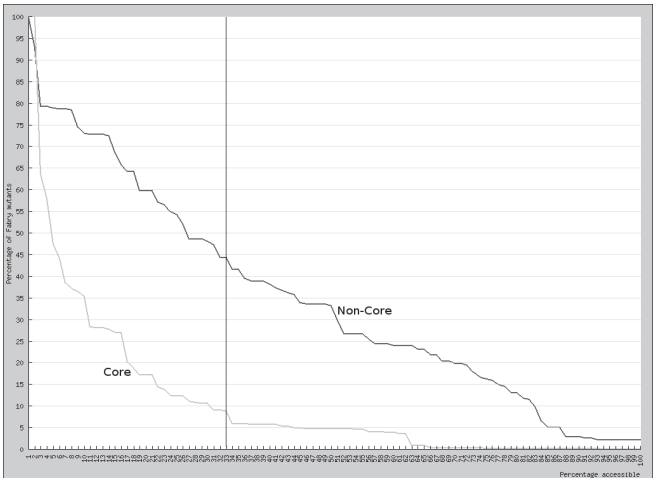


Figure 5. The correlation between Fabry disease-causing mutations at structurally conserved and non-conserved core positions and their solvent accessibility. X-axis: Percentage of accessible side chain surface area for each residue in the human GLA protein. Y-axis: Percentage of the total set of Fabry 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.

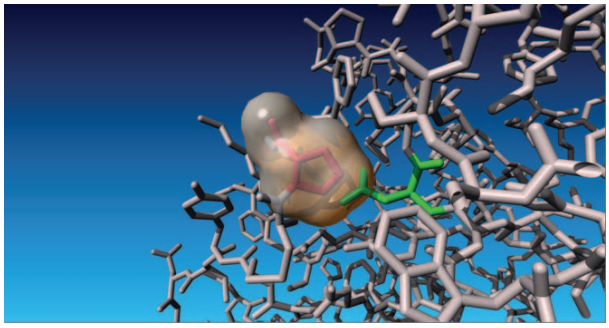


Figure 6. YASARA ball and stick backbone visualization of mutation p.A143P in GLA. The alanine to proline substitution at 3D-position 76 is depicted in red and surrounded by its Van der Waals surface. Clearly steric hindrance is observed with the side-chain of aspartate 93.

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?

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.

The α/β-Hydrolase Fold 3DM Database (ABHDB) as a Tool for Protein Engineering

Robert Kourist, Helge Jochens, Sebastian Bartsch
Remko Kuipers, Santosh Kumar Padhi, Markus Gall
Dominique Böttcher, Henk-Jan Joosten, Uwe Bornscheuer

Introduction

The α/β-hydrolase fold enzyme superfamily is one of the largest groups of structurally related enzymes [1]. Applications of the members of this versatile enzyme family range from the kinetic resolution of precursors of pharmaceutical compounds [2], degradation of pollutants[3], and bulk applications such as lipid modification [4] and laundry detergents[2a]. The family is characterized by the α/β-hydrolase fold, consisting of a central β-sheet surrounded by several a-helices [5] and by a common catalytic triad formed of a catalytic nucleophile, a histidine and an acidic residue. The superfamily covers a large diversity of catalytic activities like the hydrolysis of carboxylic acid esters, carboxylic acid amides, thioesters and epoxides, C-C bond-formation by hydroxynitrile lyases and the dehalogenation and haloperoxidation of organic compounds.[6] In addition, many α/β-hydrolase fold enzymes show 'catalytic promiscuity', e.g. they are able to catalyze more than one type of chemical transformation.[7]

Protein engineering proved to be an efficient method to tailor α/β-hydrolase fold enzymes towards a desired property, for instance with a drastically improved thermostability [8] and with completely inverted enantiopreference [9]. Moreover, enzymes with completely new catalytic activities have been generated. Impressive examples are the conversion of an esterase from *Pseudomonas fluorescens* into an epoxide hydrolase [10] a perhydrolase [11] and of an esterase [12] or a heme-free bromoperoxidase [13] into enzymes with lipase-like properties. Lipases can be engineered by implementing major structural modifications such as the exchange of secondary structure elements in lid swapping [14] or even major structural modifications such as circular permutation [15]. Nevertheless, despite an impressive progress over the last years [16], protein design has not been established as a routine procedure so far. One of the major bottlenecks lies in the insufficient knowledge about structure-function relationships and the unsatisfactory reliability of *in silico* predictions. For α/β-hydrolase fold enzymes, two major strategies have been used to overcome this limitation: Molecular modeling of the enzymatic mechanisms [3, 9d] or an analysis of structural relationships between different enzymes.[17] The ultimate objective of a computational tool for protein engineering applications would be to make reliable predictions on the outcome of amino acid exchanges on the properties and performance of a given enzyme.

Due to the limitations in predictibility, random approaches, often referred to as directed evolution [16, 18] are still the method of choice. The major advantage of directed evolution compared to all other approaches is its relative independence from information input. In principle, the gene encoding the targeted protein is sufficient to perform the experiment with a reasonable chance to find improved variants, but only if fast and reliable screening or selection systems are available.

Powered by rapidly increasing availability of information about proteins such as structures, sequences and biochemical data, and the ongoing progress in computational technologies directed evolution experiments are more and more moving towards

combination with rational concepts. The computational part of those experiments is thereby often limited to the prediction of amino acid positions that are likely to influence a desired enzyme property without suggesting specific amino exchanges. Reetz *et al.* proposed creating the diversity only in those parts of the enzyme that are believed to have an influence on the targeted enzyme function [19]. Thus, by subsequently randomizing active site residues (CASTing) in the epoxide hydrolase of *Aspergillus niger* the enantioselectivity of this enzyme could be evolved from $E = 4.6$ to 115 [18d]. This example illustrates that directed evolution is not anymore the naïve introduction of mutations followed by screening or selection, but has developed to a more focused diversification that concentrates on certain protein regions to create superior protein libraries. A recent study on the focused directed evolution of a Baeyer-Villiger monooxygenase demonstrated that high quality alignments are of particular value for the identification of potential randomization sites, especially in cases where a highly dynamic molecular mechanism makes predictions of the result of amino exchanges difficult [20].

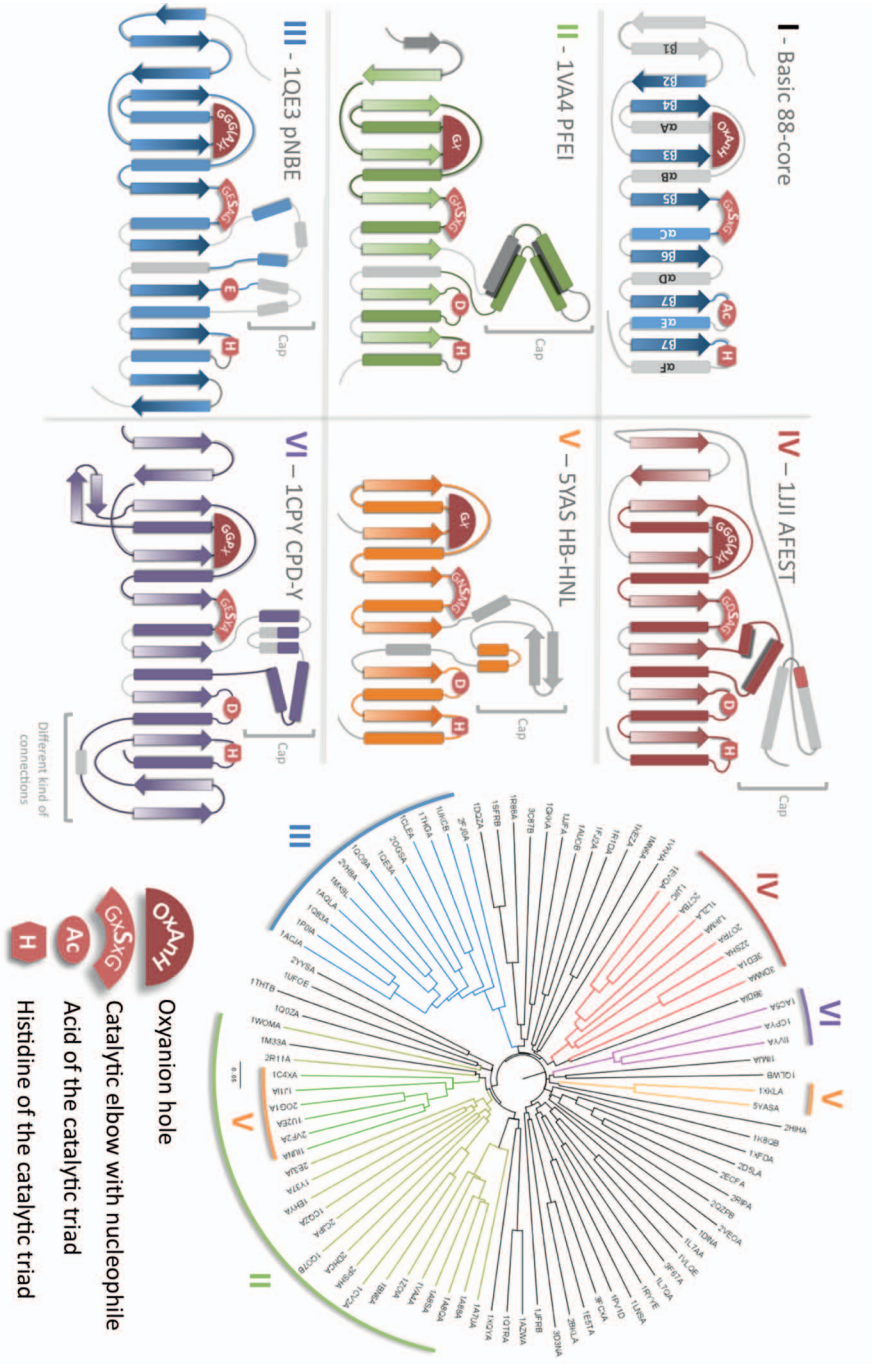
Classical molecular modeling of α/β-hydrolase fold enzymes has mainly focused on the tetrahedral intermediate, which is assumed to resemble the rate limiting step [21]. Recently, other steps than the tetrahedral intermediate came into the focus. For instance, the accessibility of the active site to the substrate [10, 22] or water access tunnels [3, 23] can also be crucial for the selectivity of an enzyme. Hult and co-workers pointed out the importance of entropy in the hydrolysis of chiral alcohols [9d, 24]. Although molecular modeling of the catalytic mechanism of hydrolases has led to multiple successful predictions, no general applicable method has been developed yet. This can in part be attributed to the limited information value of the available structural data. Most crystal structures of α/β-hydrolase fold enzymes represent empty enzymes or have an inhibitor in the active site, whereas structures with the substrate of interest are the exception. Moreover, catalytic properties, such as the enantioselectivity, are often governed by very subtle effects in the catalytic site and are difficult to predict. Furthermore, recent studies suggest that an enzyme can undergo several sub-states during catalysis, which are difficult to determine by X-ray crystallography [25]. Nevertheless, due to ongoing developments in X-ray crystallography and NMR spectroscopy it can be expected that the quality of future structural data may increase. In addition, progress in the generation of smart tools for computational biochemistry raises access to a better understanding of molecular catalysis in enzymes [26].

A stringent classification of the α/β-hydrolase fold super-family according to function, amino acid sequence and three-dimensional structure would provide the basis for a deeper analysis of structure-function relationships and would facilitate the identification of target amino acids for directed evolution. However, the classification of the vast and growing amount of sequential and structural data on α/β-hydrolase fold enzymes has been hampered by the high degree of diversity within the superfamily. Analog functional properties of its members do often not correlate with similarity of the three-dimensional structure or the amino acid sequence. For instance, lipases share common features such as activity towards

long-chain fatty acids and interfacial activation but are enzymes with a high structural and sequential diversity [17a]. Arpigny and Jaeger [17a] proposed a classification of bacterial lipases into eight classes based on their amino acid sequences and functional properties. Several databases for specific groups of α/β-hydrolase fold enzymes such as carboxylesterases [27-29], lipases [30], and epoxide hydrolases [31] have been established and reveal structural consensus motifs, which resulted in valuable insights into structure-function relationships. Fischer and Pleiss [30] suggested that most lipases and esterases can be classified according to the composition of their oxyanion hole into GX- and GGG(A)X-hydrolases. Interestingly, functional properties such as the activity of esterases toward tertiary alcohols could be later derived from this classification [17b]. In contrast to the diversity found in lipases, hydroxy nitrile lyases, epoxide hydrolases and haloalkane dehalogenases have a very high structural similarity to GX-carboxyl ester hydrolases. Nevertheless, the transformation of an esterase to an epoxide hydrolase by means of directed evolution proved to be very difficult and was only achieved by the generation of a chimeric hybrid enzyme [10]. These examples underline the challenges for a proper classification of the α/β-hydrolase fold enzyme super family.

The α/β-Hydrolase Fold Enzyme Family 3DM Database (ABHDB) is a high quality structure based multiple sequence alignment that is based on almost all available α/β-hydrolase fold enzymes and is composed of separate subfamily sequence alignments of subfamilies where a structure is available [32]. 3DM is a tool that can automatically generate super-family databases and has been applied to different protein superfamilies such as receptors, antibodies, and several enzyme protein families. These databases are used in the early stages of drug design [32, 33] for validation of undetermined variants in DNA diagnostics [34], and as guidance in enzyme engineering projects targeting different enzyme features. For instance, 3DM enabled prediction of mutations that changed enzyme specificity of an oxaloacetate hydrolase (OAH) [37], and it was used to selectively remove OAH activity from the petal death protein [35]. 3DM was also applied for the suggestion of specific mutations that increase enzyme activity of phosphoglucose isomerase [36] and the oxidation rate of a FAD-binding oxidase [37]. The automated structure of 3DM allows for easy updating of the database. This is vital in order to keep pace with the fast growing amount of data. 3DM tools, such as *Comulator*[36], *Mutator* [34], a sophisticated filter management system and a systematic numbering scheme facilitate the analysis of correlations between amino acids and the fast tracking of literature on mutagenesis data of specific positions in an α/β-hydrolase fold enzyme.

Figure 1 (page 179). Conserved regions in different groups of the 3DM-based α/β-hydrolase fold enzyme database. The colouring of the phylogenetic tree corresponds to the colours of the schematic representations of the α/β-hydrolase folds of the different families. Parts shown in grey are variable regions of the example protein, the coloured parts represent the conserved core region of the corresponding family. The following proteins were used as representative examples: 1VA4: Pseudomonas fluorescens aryl esterase (PFEI); 1QE3: Bacillus subtilis p-nitrobenzyl esterase (pNBE); 1JJI: Archaegobus fulgidus carboxylesterase (AFEST); 5YAS: Hevea brasiliensis hydroxynitrile lyase (HB-HNL); 1CPY: Saccharomyces cerevisiae carboxypeptidase Y (CPD-Y).


I - Basic 88-core
II - 1VA4 PFEI
III - 1QE3 pNBE
IV – 1JJI AFEST
V – 5YAS HB-HNL
VI – 1CPY CPD-Y
Cap
Different kind of connections
Oxyanion hole
Catalytic elbow with nucleophile
Acid of the catalytic triad
Histidine of the catalytic triad

Thus, ABHDB provides protein engineers with a practical tool for the identification of target amino acid residues in members of the α/β-hydrolase fold enzyme superfamily and allows a facilitated analysis of the results. This article intends to provide an overview on the scope of the α/β-hydrolase fold enzyme superfamily and the corresponding 3DM database and gives some recent examples for its application in protein engineering.

Aligning the haystack: Generation of a database containing most enzymes of the superfamily

The standard 3DM method with default settings [32] was used to generate the α/β-hydrolase fold enzyme superfamily database (http://fungen.wur.nl/ABHDB). However, due to the size and complexity of this super-family, 3DM needed adjustments and several extensions were generated. The first step in the generation of a 3DM database is the collection of sequences and structures. 3DM collected 668 pdb structure files which contain 1,172 separate chains that have the characteristic α/β-hydrolase fold. In the second step these 1,172 chains were superimposed and a common core of structural conserved positions was determined. The second step is the determination of the structurally conserved core of the superfamily. Using default cut-off values, 3DM detected a common core of 88 structurally conserved residue positions. In 92 of the 1,172 available structures, 3DM was not able to automatically detect the required number of core residues (default 3DM cut-off: >90%). These structures were excluded from the initial set. The third step is the determination of representative structures that will be the basis for the generation of distinct subfamilies. Normally this step is performed manually, but the size of this family requested for automation. A script was developed that chooses the representative structures automatically. To find these, 3DM uses the structural core alignment to generate a phylogenetic tree (figure 1) thereby forming groups of the sequences. Each group contains sequences that are at least 80% identical at the core positions. To perform this step 3DM uses the Levenstein algorithm [38]. After grouping all sequences, the best representative was chosen using the following criteria: (1) The number of core residues that could be detected by 3DM for a specific sequence (each sequence can maximally miss eight residues, since the cut-off for inclusion is 90% of all core residues). Obviously, sequences for which the core was completely detected are the best candidates to serve as sub-family templates; (2) The number of sequentially missing core residues: sequences for which 3DM could not detect the core for two or more neighboring residues are not taken as template; (3) The reason for missing core residues: sometimes pdb files have gaps in the structures because these parts of the structure could not be solved by X-ray analysis. In these cases 3DM excludes the files from the list of possible template structures. If a structure has a single gap in a core and the corresponding sequence of the structure contains exactly one residue at this gap position (only outside the cut-off of 2.5 Å), 3DM will simply add this residue to the core. 3DM will then select the structure for which this residue has the smallest error with respect to phy-psi angles. (4) The quality of the structure: If criteria (1) to (3) do not lead to one template candidate for a subfamily,

3DM chooses the best candidate using structure quality (e.g the resolution or RSMD [root-mean-square deviation] value) as delimiters. This procedure resulted in a core alignment that contains 99 distinct subfamily templates. In the fifth step of the database generation, superfamily sequences for which no structure is available, are collected by BLAST [39] searches using the template sequences as queries. To cope with the size of this superfamily an automated iterative BLASTing method was designed that was used to iteratively blast the sequences of these 99 templates against the Swiss-Prot and TrEMBL databases (cut-off E-value $1e^{-5}$) resulting in a set of 14,998 sequences. In the last step, 3DM aligns all resulting superfamily sequences to all templates using a four-step iterative profile based alignment procedure [40]. It then generates separate subfamilies by assigning each sequence to the nearest template using the similarities that were generated during the alignment procedure. Due to the size of the α/β-hydrolase fold enzyme superfamily iteratively aligning all 14,998 sequences to all 99 template structures would require too much CPU time and would result in large amounts of useless alignment data. Therefore, a new 3DM extension was developed that uses the BLAST results to preselect sequences that are related to starting template sequences. Manual inspection showed that BLAST cut-offs of $1e^{-90}$, $1e^{-70}$, $1e^{-30}$, and $1e^{-5}$ result in sets of sequences that can reliably be aligned in the four consecutive steps of the 3DM alignment procedure. This procedure led to the generation of 99 separate subfamily alignments. These separate subfamily alignments were combined using the core alignment of step (3) as guidance, which resulted in a superfamily alignment that contains a total of 12,431 sequences.

Structure of the α/β-hydrolase fold superfamily database

A consequence of the high diversity is that only 88 amino acids are sufficiently conserved throughout the superfamily **I** to form a consensus core. For most applications it would not be necessary to include all available structures into an alignment. The most related might be sufficient and even allow a more detailed analysis. Therefore, five subsets (**II**-**VI**) of evolutionary more closely related members were generated (figure 1, table 1). Since these members share a higher similarity, the consensus folds of these subsets contain a high number of structurally conserved residues. The phylogenetic analysis shows that these five subsets cover most of the superfamily structures. The attempt to generate a lipase-specific subset failed due to the high structural diversity of these enzymes. Figure 1 gives an overview of the numbers of subfamilies, consensus fold residues and details about the catalytic triad. 3DM designs a general numbering scheme (3D-numbers) for structural equivalent positions. This numbering scheme is applied to all data, such as sequences, structures, mutational data, co-evolution data, conservation data, amino acid contact data, etc. The systematic numbering connects all the different data types to each other which in turn enables detection of hidden correlations between different data types. Conserved residues are given with the Roman numeral of the family and the Arabic number of the residue in the consensus fold, e.g. **I_1** stands for the first residue of the superfamily **I**. The systematic 3DM numbering can be

used for all enzymes of the five groups of α/β-hydrolase fold enzymes. Hence, its systematic use would greatly facilitate the tracking of mutagenesis data on these enzymes by keyword searches in databases.

	Group	*Sub-families*	*Seqs.*	*CS.*[a]	*Var.*[b]	*Catalytic triad*[c]		
						Nucleophile	*Acidic residue*	*His*
I	Superfamily	99	12,430	88	1,468	**32** 87% S 8% D	**64** 74% D 19% E	**87** 99% H
II	GHSXGG	22	2,811	194	694	**83** 63% S 35% D	**156** 84% D 8% G	**179** 99% H
III	GESAGA	13	1,350	325	1,243	**156** 92% S 6% G	**233** 91% E 7% D	**273** 97% H
IV	GDSAGG	7	1,939	217	405	**96** 96% S	**174** 84% D 14% E	**204** 86 % H, 4% E
V	GNSMGG	7	344	180	96	**67** 96% S	**132** 97% D	**160** 98% H
VI	GESYAG	3	575	324	253	**122** 99% S	**261** 98% D	**310** 98% H

Table 1. Structure of ABHDB. a) Core size. The consensus core contains the structurally most conserved amino acids of a group of enzyme structures; b) Variability. Number of mutations within the conserved amino acids; c) The residues according to the 3DM numbering are given in bold numbers; % of the amino acid in the active site.

Overview of the α/β-hydrolase fold superfamily

Family **I** covers most of the known α/β-hydrolase fold enzymes. The high diversity of these enzymes limits the size of the conserved section to 88 amino acids. The highest degree of conservation is mostly found in defined secondary structures in the inner core of the enzyme, mainly in the central β-sheet. The peripheral regions, and, interestingly, the substrate binding sites are much less conserved. The consensus section includes β-strands 1-7 of the central β-sheet and α-helices C and E (Figure 1) of the α/β-hydrolase fold proposed by Ollis et al. in 1992 [1].

β-strand 5 and α-helix C, the so-called catalytic elbow [1], the residues of the catalytic triad with the acidic residue **I_64** and the catalytic nucleophile **I_32** are part of the consensus section. The highly conserved catalytic histidine **I_87** is present in a loop of four amino acids (**I_84-88**). The residues of the GXSXG consensus motif are highly conserved throughout the whole superfamily: Gly **I_30** 94%, Ser **I_32** 63%, Gly **I_34**

70%. Residues at positions X were assumed to be variable. For true lipases, a more specified consensus pattern [LIV]-X-[LIVFY]-[LIVMST]-G-[HYWV]-S-X-G-[GSTAC] (PROSITE pattern PS00120: LIPASE_SER) could be assigned. It should be kept in mind however, that the term true lipase refers to a class of functional similar enzymes that have very diverse structures. An analysis using 3DM shows that the variable residues X are indeed highly conserved in the individual enzyme families of the superfamily (Figure 1), which allows a division of the α/β-hydrolase fold enzyme superfamily according to the composition of the catalytic elbow into five families: **II**. GHSXGG, **III**. GESAGA, **IV**. GDSAGG, **V**. GNSMGG and **VI**. GESYAG (Table 2). These findings clearly suggest that these positions are involved in determining the specificity of the different hydrolases.

	Group	*GXSXG-lipase consensus motif*					
		G	*X*[a]	*S*[b]	*X*	*G*	*G*
I	Superfamily	**30** 94% G	**31** 25%H 16%E 12%D	**32** 87% S 8% D		**33** 93%G	**34** 70% G
II	GHSXGG	**81** 83% G	**82** 54% H	**83** 63% S 35% D	**84** >90% M or hydrophobic residue	**85** 97% G	**86** 66% G 22% A
III	GESAGA	**154** 99% G	**155** 64% E	**156** 92% S 6% G	**157**[c] 90% A	**158** 100% G	**159** 46% A 40% G
IV	GDSAGG	**94** 100% G	**95** 66% D 10% E	**96** 96% S	**97**[c] 81% A	**98** 100% G	**99** 80% G 20% A
V	GNSMGG	**65** 94% G	**66** 64% N 22% H	**67** 96% S	**68** >90% M or hydrophobic residue[c]	**69** 97% G	**70** 98% G
VI	GESYAG	**120** 95% G	**121** 90% D 7% E	**122** 99% S	**123** 96% Y	**124** 77% A 16% G	**125** 99% G

Table 2. The GXSXG-lipase pattern is highly conserved in the individual enzyme families. a) assumed to be variable in lipases; b) catalytic nucleophile; c) involved in the stabilization of the oxyanion in the catalytic mechanism; The residues according to the 3DM numbering are given in bold numbers.% of the amino acid in the active site. sidues according to the 3DM numbering are given in bold numbers; % of the amino acid in the active site.

The composition of the oxyanion hole differs between the enzyme families. It is consequently not part of the 88 amino acid core. In many cases it is preceded by the conserved histidine **I_16**. In the oxyanion hole, the consensus motifs **II**. HGX, **III**./**IV**. HGGG(A)X, **V**. HSG and **VI.** NGGP can be found (table 3). The structure of the oxyanion hole has been associated with the rare ability of some esterases and lipases to convert tertiary alcohols [17b], which has facilitated the isolation of highly enantioselective enzymes for the synthesis of these difficult substrates [41]. The knowledge about conserved structure-function relationships can thus be used to guide the search or novel enzymes.

GHSXGG: Epoxide hydrolases, dehalogenases, perhydrolases and GX-esterases

Family **II** contains 2,811 members of highly diverse functionalities including epoxide hydrolases, haloalkane dehalogenases, haloacid dehalogenases, haloperoxidases and GX-esterases. The 194 amino acids of the conserved core include the main α/β-hydrolase fold except the αD-helix of the main domain and three out of four helices that form the double V-shaped cap domain.

As expected, either a serine or an aspartate is situated on the nucleophile position **II-83** since the serine reflects esterases and haloperoxidases while the aspartate could be found mainly in haloalkane dehalogenases, haloacid dehalogenases and epoxide hydrolases (table 2). The catalytic nucleophile is most often flanked with a GHNu$^-$XG-motif. Nu$^-$ represents the nucleophile and in most cases X stands for methionine (25%) or a hydrophobic residue. The oxyanion hole of class **II** enzymes is composed of an HGX-motif. Thereby the nature of X is relatively diverse, but most often a voluminous residue such as Phe (31%) or Trp (20%) is present at this position.

Despite the high similarity between the esterases, epoxide hydrolases and dehalogenases, conserved key residues determining the chemoselectivity of the enzymes were found: C-terminal to the oxyanion hole (pos. **II_26**) Pro or Gly are highly accumulated in the alignment. Interestingly, this proline correlates to almost 100% with aspartate as catalytic nucleophile while the position is less conserved when the nucleophile is a serine. It can be assumed that this proline has a high importance for specificity of epoxide hydrolases and dehalogenases. Furthermore, there is a high correlation between the catalytic aspartate and the position **II_84**. A conserved tryptophane in this position next to the catalytic nucleophile is known to be important for the stabilization of the halide in haloalkane dehalogenases. A similar observation could be made for epoxide hydrolases: Amino acids on position **II_128** are predominantly tyrosines if the nucleophile is an aspartate and thus are believed to represent the residue involved in the protonation of the epoxide.

Group		*H/N*	*G*	*G*	*G(A)/P*	*F/W/Y/L*
I	Superfamily	**16** 65% H[c] 13% Y	n.i.[a]	n.i.[a]	n.i.[a]	n.i.[a]
II	GHSXGG	**23** 94%H	**24** 95% G	**25** 31% F 20% W	-	**25** 31% F 20% T
III	GESAGA	**81** 56% H 30% Y 8% F	**82** 95% G	**83** 98% G	**84** 64% G 21% A 10% S	**85** 64% F 18% Y 12% L 3% W
IV	GDSAGG	**31** 98% H	**32** 97% G	**33** 97% G	**34** 82% G 10% A 4% S 3% C	**35** 55% F 23% W 14% Y
V	GNSMGG	**7** 98% H	**8** 94% G 4.4% T (HNLs)[b]	**9** 41% S 27% G 13% A	**10** 71% G 19% C	
VI	GESYAG	32 65%N	33 99% G	34 99% G	35 97% P	36 96%G

Table 3. The composition of the oxyanion hole in different enzyme families. a) n.i.: not included in the consensus fold; b) hydroxynitrile lyases have a highly conserved threonine in this position; c) the residues according to the 3DM numbering are given in bold numbers. % of the amino acid in the active site.

GESAGA-family: Acetylcholine esterases, liver esterases, fungal lipases and related enzymes

The 1,350 members of this enzyme family are mostly lipases and esterases from gram positive bacteria, fungi and animals, such as acetylcholine esterase or esterases from pig liver, from *Bacillus subtilis* and *Candida rugosa* lipase. The large consensus core of 325 amino acids spans the greatest part of the central β-sheet, the surrounding α-helices, two α-helices of the cap region, the catalytic triad and the GGG(A)X motif within the oxyanion hole (figure 1) [30]. The core is longer than that of the closely related family **IV** and contains two additional strands at the C-terminus. The active site of family **III** is rather exposed to the solvent. The catalytic elbow bears a highly conserved GESAGA motif (table 2). Family **III** is the only one where the acidic residue of the catalytic triad is glutamate instead of aspartate. The C-terminal neighbor of the catalytic **III_156** serine is mostly alanine **III_157**. This residue is often involved together with residues from the GGG(A)X-motif in the formation of the oxyanion hole [30]. The N-terminal neighbor of the catalytic nucleophile is a highly

conserved (65%) glutamate **III_155**. This residue has been discussed in the context of the ability of the promiscuitiv activity of some esterases of family **III** towards carboxylic acid amides [42] and forms a highly conserved triangle of 5-7 Å with the catalytic glutamate **III_233** and the acidic residue **III_276** (63% E, 34% D). Interestingly, the three charged residues are positioned in proximity of the catalytic machinery. They are connected by hydrogen bridges via water molecules. The high degree of conservation of this structural motif suggests an important function in catalysis.

GDSAGG: Hormone-sensitive lipase-like

The 1,950 sequences of **IV** include animal lipases such as the hormone-sensitive lipase, enzymes from bacteria and from archea and a high number of enzymes from the metagenome. Several proteins without enzymatic activity such as a putative gibberellin receptor GID1L1 from *Arabidopsis thaliana* [43] are also included. Interestingly, group **IV** contains several thermophilic esterases.

The central β-sheet with the surrounding α-helices is represented in the consensus fold of 217 amino acids. Family **IV** shares with family **III** the GGG(A)X-motif in the oxyanion hole (figure 1). Several structural features distinguish both enzyme families: The acidic residue **IV_174** of the catalytic triad of family **IV** is mostly aspartate, not glutamate as in **III**. Also the N-terminal neighbor of the catalytic serine is an aspartate, and the Glu-Glu-Glu triangle of family **III** has no equivalent in **IV**. Interestingly, the N-terminus of family **IV** is part of the cap-region and shields the active site (figure 1).

GNSMGG: Hydroxynitrile lyases and related enzymes

Enzymes present in this family are mainly hydroxynitrile lyases (HNLs), esterases (salicylic acid-binding proteins) and C-C bond hydrolases, while the first two are plant enzymes; the latter are from bacterial sources. The catalytic triad is highly conserved in this family **V** (Table 1). The HNLs use two additional amino acids Thr **V_8** and Lys **V_161** for catalysis, which are conserved among only the HNL subfamily, Thr **V_8** (100%) and Lys **V_161** (92%). Their consensus is, however, low in the GNXMG-family (only 4.4% and 3.5% respectively), which clearly differentiates the diverse function of HNLs in the family. Five of seven subfamilies of the GNSMGG-family are C-C bond hydrolases, belonging to the group of hydrolases of *meta*-fission products (MFP) also referred to as *meta*-cleavage product hydrolases. These enzymes share the ability to hydrolyse products of the degradation of aromatic compounds such as catechols [6,44]. Interestingly, most of the sequences of MFP hydrolases appear in family **V** and **II**, which is also demonstrated in figure 1. The catalytic motif GXSXG, in C-C bond hydrolases is GNSM(F)GG where the methionine stabilizes the oxyanion substrate in the catalytic site. Interestingly, the motif in the esterase- (**II**) and HNL- (**IV**) subfamilies is GHSLGG and GESCG(A)G, respectively, which proves that a difference in structure/sequence relates to the function of the enzyme. Another highly conserved motif **V_7-10** (Table 2) important for catalysis despite of the dissimilar functions

of the family is observed throughout the family. It constitutes (a) a part of the oxyanion hole in case of SABP2 esterase (main chain of Ala, **V_9**) and in case of C-C bond hydrolysis reaction catalyzed by BphD [45], (b) a main catalytic residue for HNL activity, **V_8** (Thr) which stabilizes the hydroxy group of cyanohydrins by H-bonding. Like the role of the catalytic motif in other enzymes, Glu **V_66** and Cys **V_68,** the adjacent residues of catalytic nucleophile Ser **V_67** are important for HNL activity as their replacement with A drastically reduced the HNL activity mainly in *Hevea brasiliensis* HNL [46]. The cap domain of HNLs consists of three α-helices and two β-sheets. Like most other α/β-hydrolase fold enzymes the cap domain of HNLs lies in-between the nucleophile Ser and the acid Asp, but the β-sheets are not common in the superfamily.

GESYAG: Carboxypeptidases and related enzymes

Family **VI** contains carboxypeptidases and a high number of uncharacterized proteins, mainly from yeast, but also from animals and plants. Family **III** with 325 amino acids and **VI** with 324 are by far the largest ones of the α/β-hydrolase fold enzyme superfamily. The consensus core with nine α-helices and 12 β-strands is well conserved. The carboxypeptidases have an insertion of two conserved β-strands between α-helix A and β-strand 4. Interestingly, the connections of the conserved C-terminal α-helices and 12 β-strands differ from all other enzyme families. Family **VI** shows some similarity to families **III** and **IV**. In the oxyanion hole a conserved GGP-motif with some similarity to the GGG(A)X-motif can be found. The different structure of the C-terminus of family **VI** distinguishes it from the other families within the α/β-hydrolase fold superfamily.

Enzymes not covered by the classification

Several sequences and structures could not be classified into any of the homogenous enzyme families **II**-**VI**. A prominent member of these enzymes is lipase A from *Candida antarctica* [47]. This lipase is distinguished by several unique catalytic properties such as the ability to convert tertiary alcohols and a preference towards the *trans*-isomers of unsaturated fatty acids [48]. Its recently resolved structure cannot be grouped into the GX- or the GGG(A) X-classes [47]. This presumably also holds true for related lipases from *Ustilago maydis* [49] and *Kurtzmanomyces* sp. I-11 [50]. Several other bacterial and fungal lipases differ to a high degree from the other known α/β-hydrolase fold enzymes and cannot yet be classified by the structural alignment using 3DM.

ABHDB as a tool for protein engineering

The classification of esterases and lipases according to the composition of their oxyanion hole distinguishes enzymes bearing one glycine (GX) followed by a voluminous amino acid (X) from enzymes bearing three glycines or two glycines followed by an alanine (GGG(A)X, table 2) [30]. These enzymes can be found in families **III** and **IV** of the ABHDB. The consensus pattern has been related to the activity of α/β-hydrolase fold enzymes towards tertiary

alcohols [17b]. Recently, it was found that esterase EstA from *Paenibacillus barcinonensis* has a very low activity towards tertiary alcohols. It belongs to class **III** and has a serine **III_84** in the third position of the consensus motif. Serine in this position has an abundance of only 10% according to 3DM. A back mutation to the highly conserved glycine yielded an enzyme variant with strongly increased activity towards tertiary alcohols [51]. Furthermore, the position **III_82** had been shown previously to be a determinant of the enantioselectivity as a sixfold increased enantioselectivity was observed when the mutation Gly **III_82** to Ala was introduced into esterase BS2 from *B. subtilis* [41a]. Interestingly, the analog mutant **III_82**A of EstA also had a strongly increased enantioselectivity in the kinetic resolution of tertiary alcohols. This example underlines the value of homology-based analyses of structure-function relationships for the engineering of catalytic properties of enzymes.

The vast protein data set that is ordered and stored by 3DM is of high value for the guidance of directed evolution experiments. Due to the procedure of 3DM to create structure-based multiple sequence alignments, all amino acids in a given sequence within the 3DM core can be assumed to be on the same structural position in every enzyme that is included in the alignment. Thus, the program allows the determination of the amino acid distributions at these positions. Those distributions can be effectively used to generate 'smart' mutant libraries for a saturation mutagenesis. We have already successfully applied this approach to improve the enantioselectivity of an esterase from *Pseudomonas fluorescens* by variation of four residues in the acyl-binding pocket [52], resulting in the identification of variants with substantially increased enantioselectivity [53] within a set of only 500 variants screened. This success was attributed to the high quality of the libraries designed using ABHDB analysis reducing the theoretical screening effort from 3 million [54] to just 10,000 clones to cover 95% of all possible combinations. In a similar approach three surface residues of the same enzyme were targeted to increase the thermostability (unpublished).

Outlook

The vast amount of sequential and structural data on α/β-hydrolase fold enzymes is difficult to assess by standard bioinformatic methods and databases. To our knowledge, ABHDB is the first structural alignment of all known enzymes of this superfamily, which is a key feature distinguishing it from sequence-based approaches [27-30]. It represents a very helpful tool for the analysis of structure-function relationships and the mechanistic determinants of substrate specificity, even if not all known sequences could be incorporated. Despite the diverse nature of α/β-hydrolase fold enzymes it is striking that the central core of the fold and the catalytic triad is well-conserved throughout the 12,431 sequences of the database. The residues contributing to the stabilization of the oxyanion, however, differ considerably, but are itself highly conserved in different subgroups. They can be related to the functionality of the enzyme, making them useful fingerprint motifs for the classification and discovery of novel biocatalysts.

Literature search in organic chemistry is considerably easier than in protein engineering. It is very easy to confirm the state of the art on a given organic compound. A simple search for structure, sum formula or systematic name in databases like CrossFire Beilstein gives comprehensive information on synthesis, structural data and applications. This is vital to confirm the novelty of results. In contrast, there is no such a tool for mutagenesis and the scientist has to rely on his or her literature knowledge. Also a systematic nomenclature of amino acid residues in consensus scaffolds is missing. Besides qualitative designations such as 'catalytic nucleophile' or 'histidine of the catalytic triad' keyword searches in databases will not reveal specific amino acid exchanges in analog positions of different enzymes. This may lead to cases were a given amino acid exchange in an enzyme scaffold may have been investigated independently in different laboratories several times. The scientists would perhaps not be aware of each other's work due to the use of different key words or because the exchange was performed in analogue enzymes such as, for instance, a GX-esterase and a haloalkane dehalogenase. Herein we present with the 3DM numbering a systematic nomenclature for the five enzyme groups of the α/β-hydrolase fold enzymes. Usage of this system may facilitate the future tracking of mutations in these enzymes. In addition, the tool *mutator* of ABHDB connects amino acids of α/β-hydrolase fold enzymes to literature data on their mutagenesis. Despite the limitations – only conserved amino acids are analyzed and the search is confined to scientific publications but not patents – this tool would offer the technology for the future systematic assessment of mutagenesis. Such a tool would enable access to all mutagenesis done to a certain position in a scaffold. It might attain a similar significance as Scifinder and Pubmed [55] already have in the every-day life of scientists in life sciences and biotechnology.

Directed evolution has been described as finding the needle in the haystack. We believe that the ABHDB contributes considerably to ordering the vast and complex data of the α/β-hydrolase fold enzyme superfamily. This is particularly important in view of the future incorporation of vast numbers of sequences and structures from genomic and metagenomic sources that can be expected from the recent progress in sequencing techniques. The possibility of an automatic update perhaps is not the last useful tool of 3DM, as it allows keeping pace with the progress. ABHDB can be used together with molecular modelling and other computational tools to guide rational protein design. A particular interesting application of ABHDB is the generation of smart libraries of small size for directed evolution experiments – the haystack can easily be aligned before the needle gets explored.

Acknowledgements

RK and UTB are indebted to the Japanese Society for the Promotion of Science (JSPS) for stipends (P-09010, S-09200). SKP is very much grateful to the Alexander von Humboldt (AvH) foundation for his research fellowship. UTB also thanks the German Research Foundation (DFG, Grant Bo1862/4-1) and the EU (Grant PITN-GA-2007-215560 ENEFP) for financial support.

References

1. D. L. Ollis, E. Cheah, M. Cygler, B. Dijkstra, F. Frolow, S. M. Franken, M. Harel, S. J. Remington, I. Silman, J. Schrag, J. L. Sussman, K. H. G. Verschueren, A. Goldman, Prot. Eng. 1992, 5, 197-211.
2. a) U. T. Bornscheuer, R. J. Kazlauskas, Hydrolases in Organic Synthesis, 2 ed., Wiley-VCH, Weinheim, 2005; b) R. N. Patel, Curr. Opin. Drug. Discov. Devel. 2006, 9, 741-764.
3. M. Pavlova, M. Klvana, Z. Prokop, R. Chaloupkova, P. Banas, M. Otyepka, R. C. Wade, M. Tsuda, Y. Nagata, J. Damborsky, Nat. Chem. Biol. 2009, 5, 727-733.
4. U. T. Bornscheuer (Ed.), Enzymes in Lipid Modification, Wiley-VCH, Weinheim, 2000.
5. J. Pleiss, H. Scheib, R. D. Schmid, Biochimie 2000, 82, 1043-1052.
6. M. Holmquist, Curr. Protein Pept. Sci. 2000, 1, 209-235.
7. a) R. Fujii, Y. Nakagawa, J. Hiratake, A. Sogabe, K. Sakata, Prot. Eng., Des. Sel. 2005, 18, 93-101; b) E. Henke, U. T. Bornscheuer, Anal. Chem. 2002, 75, 255-260; c) K. Hult, P. Berglund, Trends Biotechnol. 2007, 25, 231-238. d) C. Li, M. Hassler, T. D. H. Bugg, ChemBioChem 2008, 9, 71-76.
8. M. T. Reetz, J. D. Carballeira, A. Vogel, Angew. Chem., Int. Ed. 2006, 45, 7745-7751; Angew. Chem. 2006, 118, 7909-7915.
9. a) S. Bartsch, R. Kourist, U. T. Bornscheuer, Angew. Chem., Int. Ed. 2008, 47, 1508-1511; Angew. Chem. Int. Ed. 2008, 120, 1531-1534; b) M. Ivancic, G. Valinger, K. Gruber, H. Schwab, J. Biotechnol. 2007, 129, 109-122. c) Y. Koga, K. Kato, H. Nakano, T. Yamane, J. Mol. Biol. 2003, 331, 585-592; d) A. O. Magnusson, M. Takwa, A. Hamberg, K. Hult, Angew. Chem., Int. Ed. 2005, 44, 4582-4585; Angew. Chem. 2005, 117, 4658-4661; e) D. X. Zha, S. Wilensek, M. Hermes, K. E. Jaeger, M. T. Reetz, Chem. Comm. 2001, 2664-2665.
10. H. Jochens, K. Stiba, C. Savile, R. Fujii, J. G. Yu, T. Gerassenkov, R. J. Kazlauskas, U. T. Bornscheuer, Angew. Chem., Int. Ed. 2009, 48, 3532-3535; Angew. Chem. 2009, 121, 3584-3587.
11. D. L. Yin, P. Bernhardt, K. L. Morley, Y. Jiang, J. D. Cheeseman, V. Purpero, J. D. Schrag, R. J. Kazlauskas, Biochemistry 2010, 49, 1931-1942.
12. D. Reyes-Duarte, J. Polaina, N. Lopez-Cortes, M. Alcalde, F. J. Plou, K. Elborough, A. Ballesteros, K. N. Timmis, P. N. Golyshin, M. Ferrer, Angew. Chem., Int. Ed. 2005, 44, 7553-7557; Angew. Chem. 2005, 117, 7725-7729.
13. B. Chen, Z. Cai, W. Wu, Y. Huang, J. Pleiss, Z. Lin, Biochemistry 2009, 48, 11496-11504.
14. a) C. C. Akoh, G. C. Lee, J. F. Shaw, Lipids 2004, 39, 513-526; b) Y. L. Boersma, T. Pijning, M. S. Bosma, A. M. van der Sloot, L. F. Godinho, M. J. Droge, R. T. Winter, G. van Pouderoyen, B. W. Dijkstra, W. J. Quax, Chem. Biol. 2008, 15, 782-789; c) M. Skjot, L. De Maria, R. Chatterjee, A. Svendsen, S. A. Patkar, P. R. Ostergaard, J. Brask, ChemBioChem 2009, 10, 520-527.
15. Z. Qian, S. Lutz, J. Am. Chem. Soc. 2005, 127, 13466-13467.
16. R. J. Kazlauskas, U. T. Bornscheuer, Nat. Chem. Biol. 2009, 5, 526-529.
17. a) J. L. Arpigny, K. E. Jaeger, Biochem. J. 1999, 343, 177-183; b) E. Henke, J. Pleiss, U. T. Bornscheuer, Angew. Chem., Int. Ed. 2002, 41, 3211-3213; Angew. Chem. 2002, 114, 3338-3341.
18. a) F. H. Arnold, Georgiou, G. (Eds.), Directed Enzyme Evolution Screening and Selection Methods, Totawa, USA, 2003; b) F. H. Arnold, Georgiou, G. (Eds.), Directed Enzyme Evolution Library Creation, Totawa, USA, 2003; c) S. Lutz, U. T. Bornscheuer (Eds.), Wiley VCH, Weinheim, 2009; d) M. T. Reetz, L. W. Wang, M. Bocola, Angew. Chem., Int. Ed. 2006, 45, 1236-1241; Angew. Chem. 2006, 118, 1258-1262.
19. M. T. Reetz, J. D. Carballeira, Nat. Protoc. 2007, 2, 891-903.
20. S. Wu, J. P. Acevedo, M. T. Reetz, Proc. Natl. Acad. Sci. U.S.A. 2010, 107 2557-2780.

21. S. Raza, L. Fransson, K. Hult, Protein Sci. 2001, 10, 329-338.
22. P. B. Juhl, P. Trodler, S. Tyagi, J. Pleiss, BMC Struct. Biol. 2009, 9, 39-40.
23. M. Wittrup Larsen, Zielinska, D.F., Martinelle, M., Hidalgo, A., Jensen, L.J., Bornscheuer, U.T., Hult, K., ChemBioChem 2010, 11, 796-801.
24. J. Ottosson, L. Fransson, K. Hult, Protein Sci. 2002, 11, 1462-1471.
25. a) E. Z. Eisenmesser, O. Millet, W. Labeikovsky, D. M. Korzhnev, M. Wolf-Watz, D. A. Bosco, J. J. Skalicky, L. E. Kay, D. Kern, Nature 2005, 438, 117-121; b) J. S. Fraser, M. W. Clarkson, S. C. Degnan, R. Erion, D. Kern, T. Alber, Nature 2009, 462, 669-673.
26. J. Damborsky, J. Brezovsky, Curr. Opin. Chem. Biol. 2009, 13, 26-34.
27. X. Cousin, T. Hotelier, P. Lievin, J. P. Toutant, A. Chatonnet, Nucleic Acids Res. 1996, 24, 132-136.
28. X. Cousin, T. Hotelier, K. Giles, P. Lievin, J. P. Toutant, A. Chatonnet, Nucleic Acids Res. 1997, 25, 143-146.
29. T. Hotelier, L. Renault, X. Cousin, V. Negre, P. Marchot, , A. Chatonnet. Nucleic Acids Res. 2004, 32, D145-D147.
30. M. Fischer, J. Pleiss, Nucleic Acid Res. 2003, 31, 319-321.
31. S. Barth, M. Fischer, R. D. Schmid, J. Pleiss, Bioinformatics 2004, 20, 2845-2847.
32. R. Kuipers, H.J. Joosten, W.J.H. van Berkel, N.G.H. Leferink, E. Rooijen, E. Ittmann, F. van Zimmeren, H. Jochens, U.T. Bornscheuer, G. Vriend, V.A.P. Martins dos Santos, P.J. Schaap, Proteins 2009, 78, 2101-2113.
33. a) S. Folkertsma, P. van Noort, J. Van Durme, H. J. Joosten, E. Bettler, W. Fleuren, L. Oliveira, F. Horn, J. de Vlieg, G. Vriend, J. Mol. Biol. 2004, 341, 321-335; b) S. Folkertsma, P. I. van Noort, R. F. J. Brandt, E. Bettler, G. Vriend, J. de Vlieg, Curr. Med. Chem. 2005, 12, 1001-1016.
34. R. K. P. Kuipers, Bergh van den, T., Joosten, H.J., Lekanne dit Deprez, R.H., Mannens, M.M.A.M., Schaap, P.J., submitted 2010.
35. H. J. Joosten, Y. Han, W. L. Niu, J. Vervoort, D. Dunaway-Mariano, P. J. Schaap, Prot. Stuct. Funct. Bioinf. 2008, 70, 157-166
36. R. K. P. Kuipers, H. J. Joosten, E. Verwiel, S. Paans, J. Akerboom, J. van der Oost, N. G. H. Leferink, W. J. H. van Berkel, G. Vriend, P. J. Schaap, Prot. Struct. Funct. Bioinf. 2009, 76, 608-616.
37. N. G. H. Leferink, M. W. Fraaije, H. J. Joosten, P. J. Schaap, A. Mattevi, W. J. H. van Berkel, J. Biol. Chem. 2009, 284, 4392-4397.
38. V. I. Levenshtein, Sov. Phys. Dokl. 1966, 163, 845-848.
39. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, J. Mol. Biol. 1990, 215, 403-410.
40. H. J. Joosten, submitted, 2010.
41. a) E. Henke, U. T. Bornscheuer, R. D. Schmid, J. Pleiss, ChemBioChem 2003, 4, 485-493; b) R. Kourist, de Maria, P. D., Bornscheuer, U.T., ChemBioChem 2008, 491-498; c) R. Kourist, S. H. Krishna, J. S. Patel, F. Bartnek, T. S. Hitchman, D. P. Weiner, U. T. Bornscheuer, Org. Biomol. Chem. 2007, 5, 3310-3313; d) R. Kourist, G. S. Nguyen, D. Strübing, D. Böttcher, K. Liebeton, C. Naumer, J. Eck, U. T. Bornscheuer, Tetrahedron: Asymmetry 2008, 19, 1839-1843.
42. R. Kourist, S. Bartsch, L. Fransson, K. Hult, U. T. Bornscheuer, ChemBioChem 2008, 9, 67-69.
43. M. Salanoubat, K. Lemcke, M. Rieger, e. al., Nature 2000, 408, 820-822.
44. a) S. Khajamohiddin, E. R. Repalle, A. B. Pinjari, M. Merrick, D. Siddavattam, Crit. Rev. Microbiol. 2008, 34, 13-31; b) C. Li, M. G. Montgomery, F. Mohammed, J. J. Li, S. P. Wood, T. D. Bugg, J. Mol. Biol. 2005, 346, 241-251.
45. S. Bhowmik, G. P. Horsman, J. T. Bolin, L. D. Eltis, J. Biol. Chem. 2007, 282, 36377-36385.

46. K. Gruber, G. Gartler, B. Krammer, H. Schwab, C. Kratky, J. Biol. Chem. 2004, 279, 20501-20510.
47. D. J. Ericsson, A. Kasrayan, P. Johanssonl, T. Bergfors, A. G. Sandström, J. E. Bäckvall, S. L. Mowbray, J. Mol. Biol. 2008, 376, 109-119.
48. P. D. de Maria, C. Carboni-Oerlemans, B. Tuin, G. Bargeman, A. van der Meer, R. van Gemert, J. Mol. Catal. B.: Enzym. 2005, 37, 36-46.
49. J. Kaemper, R. Kahmann, M. Bolker, e. al., Nature 2006, 444, 97-101.
50. K. Kakugawa, M. Shobayashi, O. Suzuki, T. Miyakawa, Biosci. Biotechnol. Biochem. 2002, 66, 1328-1336.
51. A. Bassegoda, Nguyen, G.S., Schmidt, M., Kourist, R., Diaz, P., Bornscheuer, U.T., submitted 2010.
52. S. Park, K. L. Morley, G. P. Horsman, M. Holmquist, K. Hult, R. J. Kazlauskas, Chem. Biol. 2005, 12, 45-54.
53. H. Jochens, Bornscheuer, U.T., submitted 2010.
54. M. T. Reetz, D. Kahakeaw, R. Lohmer, ChemBioChem 2008, 9, 1797-1804.
55. J. McEntyre, D. Lipman, Can. Med. Assoc. J. 2001, 164, 1317-1319.

Increasing the thermostability of sucrose phosphorylase by a combination of sequence- and structure-based mutagenesis

An Cerdobbel, Karel De Winter, Dirk Aerts
Remko Kuipers, Henk-Jan Joosten
Wim Soetaert, Tom Desmet

Abstract

Sucrose phosphorylase is a promising biocatalyst for the glycosylation of a wide variety of acceptor molecules, but its low thermostability is a serious drawback for industrial applications. In this work, the stability of the enzyme from *Bifidobacterium adolescentis* has been significantly improved by a combination of smart and rational mutagenesis. The former consists of substituting the most flexible residues with amino acids that occur more frequently at the corresponding positions in related sequences, while the latter is based on a careful inspection of the enzyme's crystal structure to promote electrostatic interactions. In that way, a variant enzyme could be created that contains 6 mutations and whose half-life at the industrially relevant temperature of 60 °C has more than doubled compared to the wild-type enzyme. An increased stability in the presence of organic co-solvents could also be observed, although these effects were most noticeable at low temperatures.

Introduction

Sucrose phosphorylase (SP) catalyzes the reversible phosphorolysis of sucrose into α-D-glucose-1-phosphate and fructose. Although it is formally classified as a glycosyl transferase (EC 2.4.1.7), the enzyme belongs to glycoside hydrolase family 13 and follows the typical double displacement mechanism of retaining glycosidases [1]. Thanks to its broad acceptor specificity, SP can be employed for the transfer of glucose to a wide variety of carbohydrates as well as non-carbohydrate molecules [2]. For example, an exceptionally efficient process for the regioselective glucosylation of glycerol has recently been developed [3]. The product is a moisturizing agent for cosmetics and is commercially available under the tradename Glycoin.

For industrial application, carbohydrate conversions are preferably performed at 60 °C or higher, mainly to avoid microbial contamination [4]. Unfortunately, SP has so far been identified only in a relatively small number of micro-organisms, none of which are thermophilic [2]. The low thermostability of available SP enzymes thus forms a serious limitation for their commercial exploitation. We have recently shown that the thermostability of the SP from *Bifidobacterium adolescentis* can be dramatically improved by immobilization, either by covalent coupling to a Sepabeads enzyme carrier or in the form of a cross-linked enzyme aggregate [5,6]. Although immobilization is often employed in industrial processes to enable recovery of enzymes, the procedure can be time-consuming and expensive, especially when carriers are involved. The creation of stable enzyme variants thus still is an attractive alternative. Furthermore, the resulting proteins could serve as robust templates for the engineering of the specificity of SP, as stable enzymes have been shown to be more tolerant towards the introduction of amino acid substitutions [7,8].

The thermostability of a protein can be increased by either random or site-specific mutagenesis [9]. The SP from *Streptococcus mutans*, for example, has been stabilized by the introduction of 8 random mutations, identified through error-prone PCR [10]. However, the resulting enzyme variant retains only 60% of its activity after 20 minutes incubation at 60 °C, which is not enough for industrial applications. The crystal structure of the SP from *B. adolescentis* has been determined [11] and can be used to select specific sites for mutagenesis in this particular enzyme. Although no general rules for the rational engineering of thermostability have yet been formulated, some trends have emerged [12]. These include introducing Pro residues [13], removing Gly, Asn and Gln residues [14], extending the electrostatic interactions at the protein surface [15], and stabilizing the dipole moment of α-helices [16].

Semi-rational approaches to the engineering of thermostability have also been developed. For example, in the so-called B-Factor Iterative Test (B-FIT) positions that display the highest flexibility (highest B-factor) in an enzyme's crystal structure are submitted to site-saturation mutagenesis [17]. The best hit obtained at one site is subsequently used as a template for randomization at another site and the process is repeated iteratively until

the desired improvement is achieved. In this way, the T_{50} of the lipase from *Bacillus subtilis* could be increased from 48 °C to 93 °C by the substitution of only 7 amino acid residues [17]. To reduce the size of the libraries and thus the screening effort, an extension of the B-FIT procedure has recently been proposed, in which the randomization at each site is limited to amino acids that are frequently present in an alignment of related sequences [18]. These residues are assumed to be more favorable for the enzyme's stability and/or activity, as they have been propagated through natural selection. This 'smart library' approach has been successfully applied to the engineering of the esterase from *Pseudomonas fluorescens*, in which the number of colonies that needed to be screened could be reduced with a factor 300 [18].

Limiting the number of variants is especially convenient for SP, because the thermostability of this enzyme depends strongly on the protein concentration [6]. The high-throughput screening of mutant libraries is, therefore, not an attractive option, as this would be complicated by the detection of false positives whose thermostability is overestimated due to their low concentration. Indeed, variant enzymes are frequently expressed at a lower level than the wild-type enzyme, even when just a single amino acid has been mutated [19]. This problem can be circumvented by creating a small number of site-directed variants that can be processed manually to eliminate the effect of protein concentration. In that respect, the proposed strategy somewhat resembles the consensus approach [20], although more than one amino acid can still be introduced at any given position.

Here, we present the engineering of the SP from *B. adolescentis* by a combination of smart and purely rational mutagenesis for increased thermostability. In addition, the solvent stability of the enzyme variants has also been determined, as this feature is often correlated with thermal stability [21,22]. From a practical perspective, the addition of organic co-solvents to the reaction mixture should enable the use of hydrophobic molecules as acceptors in the glycosylation reaction catalyzed by SP.

Materials and Methods

Design and analysis of mutations

The B-factors of all the amino acids in the crystal structure of the SP from *B. adolescentis* were extracted from PDB-file 1r7a [11] with the computer program B-FITTER (Reetz *et al.*, 2006). The effect of introduced mutations was simulated by generating a homology model of each variant with the program YASARA [23], using AMBER03 as force field for energy minimizations [24].

Site-directed mutagenesis

The gene coding for the SP from *B. adolescentis* was cloned into the constitutive expression vector pCXP14h as described previously [25]. Site-directed mutations were then introduced

with a two-step PCR protocol, using the mutagenic primers listed in the supplementary data (Table S1). PCR cycling conditions were as follows: 95 °C (3 min); 30 cycles of 95 °C (1 min), 55 °C (1 min) and 65 °C (9 min). Each reaction contained 75 ng of plasmid DNA, 2.5 U *Pfu*Ultra DNA polymerase (Stratagene), 10x *Pfu*Ultra HF AD buffer, 0.2 mM of dNTP mix (Westburg) and 0.4 µM of mutagenic primer (Sigma) in a total volume of 25 µl. After the reaction, template DNA was digested with 10 U of *Dpn*I (New England Biolabs) at 37 °C for 6 h. The PCR mixture was transformed into electrocompetent *E. coli* BL21(DE3) cells (Novagen) and the transformation mixture was plated on LB medium supplemented with ampicillin. The presence of mutations was confirmed by sequencing (LGC Genomics) of plasmids isolated with a Miniprep Spin kit (Qiagen).

Enzyme production and purification

All enzymes were produced in 1l shake flasks containing 200 mL LB medium supplemented with 0.1 g/l ampicillin. Cultures were grown at 37 °C and 200 rpm for 8 hours, after which the cells were harvested by centrifugation (6,000 rcf, 4 °C, 20 min). The obtained pellets were frozen at -20 °C until further use. The recombinant enzyme was extracted by suspending the pellet in lysis solution (50 mM NaH_2PO_4, 300 mM NaCl, 10 mM imidazole, 1 mg/ml lysozyme and 0.1 mM PMSF, pH 8) at 4 °C for 30 min, followed by sonication for 2 x 2 min (Branson 250 Sonifier, level 3, 50 % duty cycle). Cell debris was removed by centrifugation (18,000 rcf, 4 °C, 30 min). The His_6-tagged protein was purified by Ni-NTA chromatography as described by the supplier (Qiagen), after which the buffer was exchanged to 100 mM phosphate pH 7 in a Centricon YM-30 (Millipore). Protein concentrations were measured according to the Lowry method [26] using BSA as standard.

Activity assays

The phosphorolysis of sucrose was measured continuously with an enzymatic assay, in which the production of α-D-glucose-1-phosphate is coupled to the reduction of NAD^+ in the presence of phosphoglucomutase (PGM) and glucose-6-phosphate dehydrogenase (G6P-DH) [27,28]. The assay solution consisted of 2 mM EDTA, 10 mM $MgSO_4$, 2 mM β-NAD, 10 µM glucose-1,6-diphosphate, 1.2 U PGM, 1.2 U G6P-DH and 100 mM sucrose in 100 mM phosphate buffer at pH 7 and 37 °C. The absorbance was measured at 340 nm in a microplate reader 680XR (Bio-Rad). One unit (U) of SP activity corresponds to the release of 1 µmole product per minute under these conditions. All assays were performed in triplicate and had a CV of less than 10 %. The kinetic parameters were determined at pH 7 and 60 °C by measuring the initial rate of fructose release with the BCA assay [6], and were calculated by non-linear regression of the Michaelis-Menten curve with the program SigmaPlot.

Stability assays

The thermostability of the variants containing a single mutation was evaluated by incubating eppendorfs containing 100 µl of purified enzyme (diluted to 8.5 µg/ml in 100 mM phosphate buffer pH 7) for 24 hours in a water bath at 60 °C. Their residual activity was then measured and compared to the activity of the untreated enzymes. Subsequently, the beneficial mutations were combined and the resulting enzyme variants examined in more detail. Their half-life at 60 °C was determined by sampling at regular intervals until the residual activity had dropped to 50 %. The corresponding t_{50}-values were calculated from a first-order fit of the stability curves. In addition, the thermodynamic stability of the variants was determined by incubating PCR tubes containing 100 µl of purified enzyme (diluted to 5 µg/ml in 100 mM phosphate buffer pH 7) for one hour in a Gradient Thermocycler (Biometra) at 58-70 °C. Their residual activity was then measured and compared to the activity of the untreated enzymes. The T_{50}-values were calculated by fitting of the linear part of the stability curves.

The solvent stability of the enzymes was determined by measuring their activity in 0-55 % (v/v) dimethylsulfoxide (DMSO). Enzyme reactions were performed at 37, 60 and 65 °C in eppendorfs containing 200 µl of purified enzyme (diluted to 0.5 µg/ml in 150 mM phosphate buffer pH 7) and 400 µl of a water/DMSO mixture containing 75 mM sucrose. The SP activity was determined with the BCA assay and the C_{50}-values were calculated by fitting of the linear part of the stability curves.

The difference in residual activity between the wild-type enzyme and the final variant at the t_{50}, T_{50} and C_{50} values were shown to be significant by means of an unpaired, two-sided t-test ($p < 0.01$).

Results and Discussion

Increasing the thermostability by 'smart' mutations

The residues with the highest flexibility in the SP from *B. adolescentis* have been identified based on their B-factor in the enzyme's crystal structure [11]. Since SP forms a homodimer, the average B-factor of both chains was used as parameter to select the target residues (Supplementary data, Table S2). Although some variation in B-factors could be observed between the monomers, the top 6 positions consistently ranked among the most flexible residues. Interestingly, the bias towards arginine and lysine residues reported in previous studies [18,29] was hardly observed in our case, and a cluster of three aspartate residues at positions 445-447 was found to be the most flexible region. This cluster is located in a loop segment of the C-terminal domain, far away from both the active site and dimer interface.

The top 10 positions were then examined in the 3DM-database [30], which contains a structure-based alignment of all sequences (~ 5,092) in glycosidase family 13, also known as the α-amylase family. Only 2 of the target positions turned out to be structurally conserved

and thus to be represented in the alignment, *i.e.* D445 and D446. At those 2 positions, the amino acid distribution was determined within the group of SP enzymes (subfamily GH13_18), which comprises 149 sequences with an identity of at least 30 % [31]. At the former position, Pro (22.8 %) occurs more frequently than Asp (20.8 %), while both Gly (23.5 %) and Thr (17.4 %) are more abundant than Asp (16.8 %) at the latter. Pro also occurs at the latter position (4.7 %) and has been included in our library because this amino acid is known to be important for protein stability. Combinations of all the corresponding substitutions were introduced in the SP from *B. adolescentis* by site-directed mutagenesis (Table 1). The resulting enzyme variants were purified by His-tag chromatography and diluted to the same protein concentration to determine their thermostability.

Enzyme	Mutations	Specific Activity	Residual Activity	Thermostability
		(U/mg)[a]	(%)[b]	(%)[c]
wild-type	-	213.0 ± 1.9	30.0 ± 1.7	100
Variant A	D445P	204.3 ± 2.1	31.5 ± 1.2	105
Variant B	D446G	214.9 ± 3.6	29.4 ± 1.6	98
Variant C	D446T	174.0 ± 4.5	29.1 ± 1.1	97
Variant D	D446P	218.4 ± 1.4	30.0 ± 1.2	100
Variant E	D445P/D446G	202.8 ± 1.9	41.1 ± 2.0	137
Variant F	D445P/D446T	197.1 ± 4.6	41.7 ± 1.8	139
Variant G	D445P/D446P	202.5 ± 2.7	30.0 ± 1.7	100
Variant H	L306H	215.6 ± 3.1	30.3 ± 1.4	101
Variant I	Q331E	208.7 ± 4.6	42.9 ± 2.5	143
Variant J	N414D	198.6 ± 2.8	15.3 ± 1.5	51
Variant K	A498H	188.6 ± 5.2	26.7 ± 1.3	89
Variant L	Q460E/E485H	192.7 ± 3.4	42.3 ± 1.9	141
Variant M	N325D/V473H	189.2 ± 2.2	24.3 ± 1.3	81
Variant N	R393N	207.8 ± 2.7	40.8 ± 2.1	136

Table 1: Thermostability of the variants create by (semi-)rational design. a) At pH 7 and 37°C. b) After 24h incubation at 60°C. c) Relative to the wild-type.

Two SP variants showed a significantly increased activity after one day incubation at 60 °C. Indeed, the substitution of the DD-motif with either PT or PG resulted in a relative thermostability of 137 % and 139 %, respectively, compared to the wild-type enzyme (Table 1). A proline at position 445 thus seems to be crucial for thermostability, at least in combination with a threonine or glycine at position 446. In accordance with its lower amino acid occurrence, a proline at position 446 does not have a significant effect on stability. In

fact, the stabilized variants contain the most frequent residues at both positions, exactly as would be predicted by the consensus concept [20]. None of the individual substitutions are able to increase the stability of SP when the neighbouring residue is not mutated, which points towards a synergistic effect between positions 445 and 446.

Increasing the thermostability by rational design

In parallel with the 'smart' mutagenesis, additional targets for mutagenesis have been selected on a purely rational basis. To that end, the enzyme's crystal structure has been carefully inspected for amino acid substitutions that could extend the network of electrostatic interactions on its surface (variants H-M, Table 1). In addition, one mutation (variant N) has been included that could possibly increase the interaction between the enzyme's subunits. The corresponding variants were then created by site-directed mutagenesis, purified by His-tag chromatography and diluted to the same protein concentration to determine their thermostability.

Three out of the seven variants showed a significantly increased residual activity after one day incubation at 60 °C (Table 1). Two of these were designed to introduce additional salt bridges, *i.e.* Q331E and Q460E/E485H. The substitution of Q331 by a glutamate was envisaged to promote an interaction with H323, which indeed seems very plausible based on the modelled structure of the resulting enzyme variant (Figure 1A). In contrast, variant L was constructed to influence more than one interaction. Residue Q460 is located close to R455, with which a negatively charged residue should be able to interact more favourably (Figure 1B). However, the glutamate at position 485 could then perhaps cause electrostatic repulsion of E460 as well as of D490. It has, therefore, been substituted by a histidine to create a network of ion pairs. Finally, variant N was created to promote the interaction between the enzyme's subunits. Indeed, residue R393 is located at the enzyme's dimer interface and was believed to induce electrostatic repulsion by interacting with the same residue from the other monomer (Figure 1C). Replacing it with an asparagine seems to have alleviated this problem.

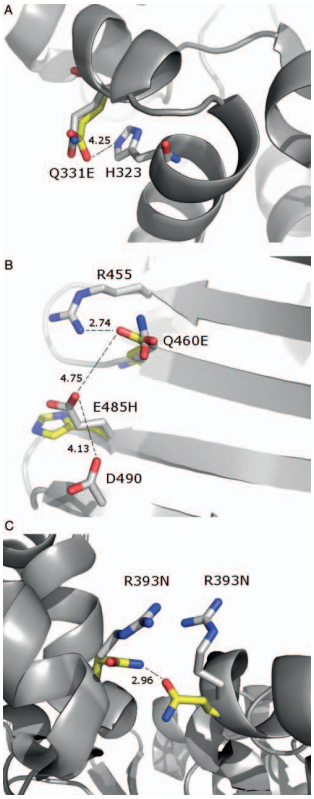


Figure 1. The rational mutations that increase the stability of SP. (A) variant I, (B) variant L, and (C) variant N. The mutations (in yellow) have been simulated in a homology model, which was superposed on the wild-type structure (in grey).

In contrast, the other four variants created by rational design did not display an increased thermostability (Table 1). These negative results are difficult to interpret, although some suggestions can be formulated based on an investigation of the variants' structural models (supplementary data, Figure S1). In variant H, for example, the substitution of L306 by a histidine was expected to introduce a salt bridge with D303. The latter might, however, already be involved in a hydrogen bond with S305 and thus not carry the required charge. In turn, the asparagine at position 414 might be crucial for the stability of SP because of its interaction with a backbone carbonyl group. This could explain why its substitution by Asp lowers the stability of variant J, although a salt bridge with R417 should have been created. Finally, the histidine residue introduced at position 498 and 473 of variant K and M, respectively, does not seem to adopt the required orientation to form a salt bridge with the neighbouring carboxylic group (D495 and D325, respectively).

Enzyme	t_{50} (h)[a]	T_{50} (°C)[b]	C_{50} (%)[c]	K_M (mM)[d]	k_{cat} (s^{-1})[d]
wild-type	24	64	34	6.5	201
variant LNFI	62	67	41	6.7	204

Table 2. Stability and activity of the optimized SP variant. a) At 60°C. b) After 1h incubation. c) Of DMSO at 37°C. d) At 60°C and pH 7.

Combining the beneficial mutations

In a next step, the different mutations were combined to examine whether they have an additive effect on thermostability. To that end, the half-life (t_{50}) of the combined variants was determined at the industrially relevant temperature of 60 °C (Figure 2). A stepwise increase in stability could be observed, resulting in a final variant (LNFI) with a half-life that has more than doubled compared to the wild-type enzyme (Table 2). Furthermore, it was found that the increased stability does not come at the expense of activity, since the k_{cat}

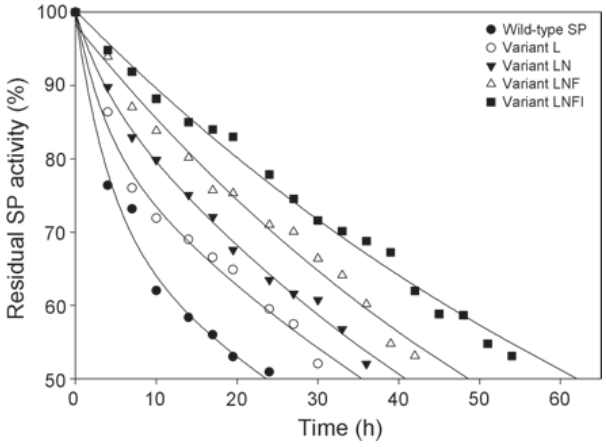


Figure 2. The kinetic stability of the improved SP variants. The enzymes were incubated at 60 °C for various times, after which their residual activity was compared with that of the untreated enzymes. All assays were performed in triplicate and had a CV of less than 10 %.

has remained constant at a value of about 200 s^{-1} (Table 2). In fact, none of the introduced mutations have a dramatic impact on the enzyme's activity, probably because the selected positions are located far away from the active site (Table 1). These results demonstrate the practical usefulness of our engineering efforts, which should stimulate the development of the various glycosylation reactions catalyzed by SP into commercial processes.

Further proof of the increased thermostability of the SP variants was obtained by determining their T_{50}-value, which is the temperature required to reduce the initial activity by 50 % after one hour incubation. This parameter reflects an enzyme's thermodynamic stability, as opposed to the kinetic stability described by its half-life [17]. The introduction of each additional mutation increased the T_{50} with nearly 1 °C (Figure 3), and the variant that contains all of the beneficial mutations displayed a T_{50} that is 3 °C higher than that of the wild-type enzyme (Table 2). This feature would probably allow the use of the SP variant

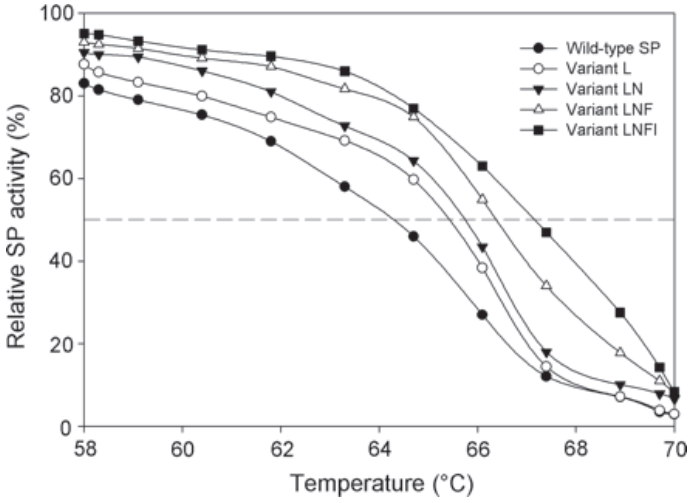


Figure 3. The thermodynamic stability of the improved SP variants. The enzymes were incubated for one hour at various temperatures, after which their residual activity was compared with that of the untreated enzymes. All assays were performed in triplicate and had a CV of less than 10 %.

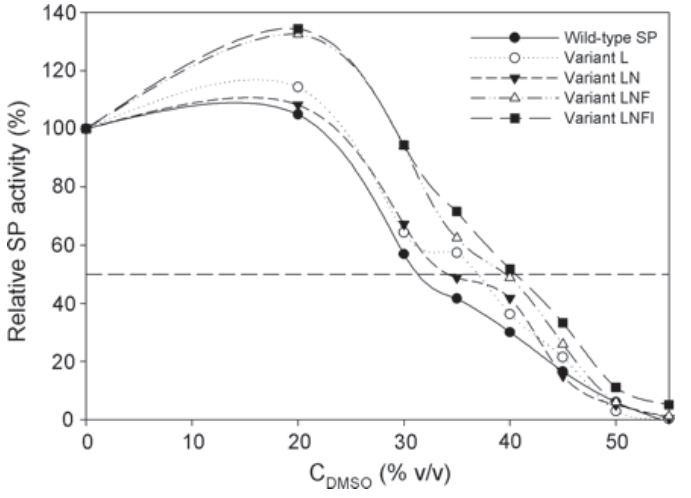


Figure 4. The solvent stability of the improved SP variants. Relative activities at 37 °C were determined in the presence of various amounts of DMSO as co-solvent. All assays were performed in triplicate and had a CV of less than 10 %.

at temperatures higher than 60 °C, although that was not the goal of the current research project.

Stability against the presence of organic co-solvents

When SP is to be applied for the glycosylation of hydrophobic molecules, organic co-solvents will need to be added to the reaction medium to increase the solubility of the acceptors. As thermostable enzymes often display a high tolerance towards other denaturing factors [21,22], the stability of the variant enzymes against the presence of DMSO as organic co-solvent has also been examined. For practical reasons, the tests were first performed at 37 °C. At that temperature, a significant improvement in solvent stability could be observed (Figure 4). Indeed, the C_{50}-value, which is the solvent concentration at which the enzyme retains half of its initial activity, has shifted from 34 % for the wild-type enzyme to 41 % for variant LNFI (Table 2).

Remarkably, an increased SP activity was observed at a DMSO concentration of about 20 %, which became even more pronounced in the stabilized variants (Figure 4). The assay temperature of 37 °C is, however, far below the optimum of 58 °C for the wild-type SP [6]. This could mean that the enzymes are too rigid for optimal functioning, a problem that might be compensated for by the addition of a low concentration of solvent. To test this hypothesis, the experiments have been repeated at higher temperatures. The 'bump' in activity was found to slowly disappear in function of temperature, being all but gone at 65 °C (Figure 5). Unfortunately, the difference in stability between the variant and wild-type SP is no longer apparent in that case, as the C_{50}-value has dropped to about 23 % for all enzymes (not shown). The thermodynamic stability of SP can thus be evidenced either by increasing the temperature or by the addition of co-solvents, but not by a combination of both.

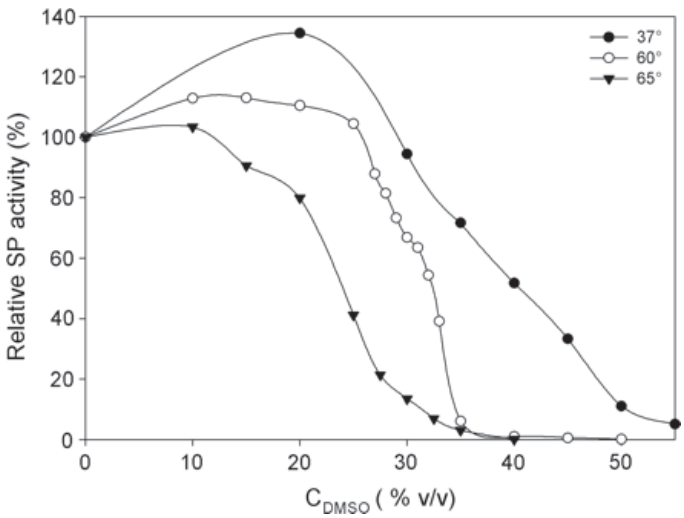


Figure 5. The influence of temperature on the solvent stability of variant LNFI. The activity of the enzyme was determined at various temperatures in the presence of various amounts of DMSO as co-solvent. All assays were performed in triplicate and had a CV of less than 10 %.

Conclusions

Sucrose phosphorylase is a promising biocatalyst for the glycosylation of a wide variety of acceptor molecules, but its use in industrial processes has been hampered by the low thermostability of the available enzymes. We have shown here that the stability of the SP from *B. adolescentis* at the process temperature of 60 °C can be efficiently improved by a combination of smart (sequence-based) and rational (structure-based) mutagenesis. In total, fourteen enzyme variants have been created, of which five displayed a considerable increase in thermostability. As expected, the introduction of consensus residues at the most flexible positions (445-446) had a positive effect on stability, although a pairwise mutation was required to achieve a synergistic effect. The rational design of thermostable variants, in contrast, proved to be a more challenging task, mainly because the interactions of the introduced amino acids with their neighbouring residues are difficult to predict. Nevertheless, creating additional salt bridges at the protein surface was found to be a successful strategy in about half of the cases.

Combining all of the beneficial mutations in a single sequence generated a biocatalyst with a half-life at 60 °C of 62 h, which is more than twice as long as the wild-type enzyme. Furthermore, the thermodynamic stability of the improved variant was also markedly enhanced, as illustrated by an increase in T_{50} from 64 to 67 °C. The addition of low concentrations (~20 %) of organic co-solvent was found to mimic the increased temperature optimum of the enzymes, but an improved solvent stability could only be observed at low temperatures. Overall, these new properties can be expected to stimulate the industrial exploitation of the various glycosylation reactions catalyzed by SP.

Acknowledgments

This work was supported by the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen) [grant number 50191].

References

1. Goedl, C., Schwarz, A., Mueller, M., Brecker, L. and Nidetzky, B. (2008b) *Carbohydr. Res.*, **343**, 2032-40.
2. Goedl, C., Sawangwan, T., Wildberger, P. and Nidetzky, B. (2010) *Biocatal. Biotransformation*, **28**, 10-21.
3. Goedl, C., Sawangwan, T., Mueller, M., Schwarz, A. and Nidetzky, B. (2008a) *Angew. Chem. Int. Ed.*, **47**, 10086-10089.
4. Haki, G.D. and Rakshit, S.K. (2003) *Bioresource Technol.*, **89**, 17-34.
5. Cerdobbel, A., De Winter, K., Desmet, T. and Soetaert, W. (2010a) *Biotechnol. J.*, **5**, 1192-1197.
6. Cerdobbel, A., Desmet, T., De Winter, K., Maertens, J. and Soetaert, W. (2010b) *J. Biotechnol.*, **150**, 125-130.
7. Bloom, J.D. and Arnold, F.H. (2009) *P. Natl. Acad. Sci. USA*, **106**, 9995-10000.
8. Tokuriki, N. and Tawfik, D.S. (2009b) *Curr. Opin. Struc. Biol.*, **19**, 596-604.
9 Lehmann, M. and Wyss, M. (2001) *Curr. Opin. Biotechnol.*, **12**, 371-375.
10. Fujii, K., Liboshi, M., Yanase, M., Takaha, T. and Kuriki, T. (2006) *J. Appl. Glycosci.*, **53**, 91-97.
11. Sprogoe, D., van den Broek, L.A.M., Mirza, O., Kastrup, J.S., Voragen, A.G.J., Gajhede, M. and Skov, L.K. (2004) *Biochemistry*, **43**, 1156-1162.
12. Eijsink, V.G.H., Bjork, A., Gaseidnes, S., Sirevag, R., Synstad, B., van den Burg, B. and Vriend, G. (2004) *J. Biotechnol.*, **113**, 105-120.
13. Matthews, B.W., Nicholson, H. and Becktel, W.J. (1987) *P. Natl. Acad. Sci. USA*, **84**, 6663-6667.
14. Kaneko, H., Minagawa, H. and Shimada, J. (2005) *Biotechnol. Lett.*, **27**, 1777-1784.
15. Serrano, L., Horovitz, A., Avron, B., Bycroft, M. and Fersht, A.R. (1990) *Biochemistry*, **29**, 9343-9352.
16. Nicholson, H., Becktel, W.J. and Matthews, B.W. (1988) *Nature*, **336**, 651-656.
17. Reetz, M.T., D Carballeira, J. and Vogel, A. (2006) *Angew. Chem. Int. Ed.*, **45**, 7745-7751.
18. Jochens, H., Aerts, D. and Bornscheuer, U.T. (2010) *Protein Eng. Des. Sel.*, **23**, 903-909.
19. Tokuriki, N. and Tawfik, D.S. (2009a) *Nature*, **459**, 668-U71.
20. Lehmann, M., Pasamontes, L., Lassen, S.F. and Wyss, M. (2000) *Biochim. Bioph. Acta*, **1543**, 408-415.
21. Cowan, D.A. (1997) *Comp. Biochem. Phys. A*, **118**, 429-438.
22. Vazquez-Figueroa, E., Yeh, V., Broering, J.M., Chaparro-Riggers, J.F. and Bommarius, A.S. (2008) *Protein Eng. Des. Sel.*, **21**, 673-680.
23. Krieger, E., Koraimann, G. and Vriend, G. (2002) *Proteins*, **47**, 393-402.
24. Wang, J.M., Cieplak, P. and Kollman, P.A. (2000) *J. Comput. Chem.*, **21**, 1049-1074.
25. Aerts, D., Verhaeghe, T., De Mey, M., Desmet, T. and Soetaert, W. (2010) *Eng. Life Sci.*, **11**, 10-19.
26. Lowry, O.H., Rosebrough, N.J., Farr, A.L. and Randall, R.J. (1951) *J. Biol. Chem.*, **193**, 265-275.
27. Silverstein, R., Voet, J., Reed, D. and Abeles, R.H. (1967) *J. Biol. Chem.*, **242**, 1338-1346.
28. Weinhausel, A., Griessler, R., Krebs, A., Zipper, P., Haltrich, D., Kulbe, K.D. and Nidetzky, B. (1997) *Biochem. J.*, **326**, 773-783.
29. Reetz, M.T., Kahakeaw, D. and Lohmer, R. (2008) *Chembiochem*, **9**, 1797-1804.

30. Kuipers, R.K., Joosten, H.J., van Berkel, W.J.H., Leferink, N.G.H., Rooijen, E., Ittmann, E., van Zimmeren, F., Jochens, H., Bornscheuer, U., Vriend, G., dos Santos, V. and Schaap, P.J. (2010) *Proteins: Struct. Funct. Bioinf.*, **78**, 2101-2113.
31. Stam, M.R., Danchin, E.G.J., Rancurel, C., Coutinho, P.M. and Henrissat, B. (2006) *Protein Eng. Des. Sel.*, **19**, 555-562.

General Discussion

Background

Detailed knowledge of all aspects of protein structure, function, and dynamics are important for many, often commercially interesting, research areas in the life sciences. In protein engineering this knowledge is applied in, for example, food and feed optimization [1], improvements of laundry detergents [2], and antibody humanization [3]. DNA diagnostics projects focus on using mutations in proteins to diagnose patients, and drug design projects study the mechanisms of proteins to create new drugs. Traditionally, these projects focused on the role of one component in one system eg. the characterization of one protein in one organism. In many such projects however, the focus is shifting to a more holistic approach by studying many proteins with similar functionality in many systems. Such systems biology like projects are however hardly possible without high-quality data sets and the proper bioinformatics tools to handle these data sets. With the ongoing advancement of high throughput acquisition technologies more and more of these data sets are becoming freely available online as producing these sets becomes both cheaper and easier. These data sets are produced using different equipment, standards, and methodologies and published using different methods and data formats. Obtaining and integrating the data required to study complex system in the life-sciences thus remains a challenge. The 3DM platform was specifically designed to retrieve and interlink available data sets in various formats for a protein superfamily and to make these data sets easily accessible for the user.

Extracting data from literature

One of the main focal points of this thesis is the application of 3DM systems in the fields of DNA diagnostics and protein engineering. In these fields a large amount of published data is available because no matter how niche the topic may be, chances are that more articles have been published about it than can be read in a lifetime. In some fields even the number of reviews is too large to keep up with. Additionally, the holistic protein superfamily based approach used by 3DM means that data obtained for related proteins are just as valuable as data obtained for the protein of interest itself. Lots of articles might therefore contain data relevant for a given study. Text mining [6] is the obvious solution to find these relevant articles and extract data from them. Many text-mining tools are available for one or more parts of the retrieval process. Projects like Open PHACTS [7] aims to use these tools to for a specific field automatically extract knowledge from the literature.

Extracting knowledge from articles is however still hampered by many challenges. The main problem is the widely varying and sometimes obfuscated language used to describe concepts and results. For Mutator, the 3DM mutation extraction tool described in chapter 5, we have therefore initially focused on retrieving only mutations from articles. There are two main reasons for selecting mutations as target. Firstly, mutations are often written relatively concise. The most common format is <wildtype><position><mutant> eg. A23W. Small variations in the way an amino acid is described and whether spaces, dashes, or other

characters are used to separate the amino acids from the position can be relatively easily included in the parser. Secondly, mutations specifically described in the text are mentioned for a reason and as such are an important source of information that can be directly used for various analyses. For example, if a mutation is retrieved from literature that has a detrimental effect on the structure or function of a non-human protein the same mutation for the human variant can most likely be classified as pathogenic.

Even such a simple format as a mutation presents various problems when extracting data from the literature. Instead of the short form a longer form may be used such as 'the alanine on position 23 was mutated to a cysteine'. Variation in such sentences is much more diverse and much harder to capture using bioinformatics tools. Another problem results from ambiguous strings such as molecular formula, gene abbreviations, or even bookmarks that might match the mutation format. This inability to reliably extract knowledge stored in articles means that much data, once published, is essentially lost. Finding all articles containing relevant data for a topic is a nontrivial task, retrieving these articles and extracting knowledge from them is even harder. Much effort is therefore duplicated and money is wasted simply because the article containing data of interest was not found or the specific data could not be automatically extracted from it. In the future we however want to expand the range of data extracted from articles beyond mutations. Due to the enormous diversity in language and writing styles found in articles it is however extremely difficult to extract specific data using software tools. There are two possible solutions for this problem. On the one hand we can continue to attempt to create more advanced and more robust text-mining tools that are capable of retrieving ever more complicated data and knowledge from articles. The alternative approach is to change the publication process itself. Guidelines already exist for sequences and structures in publications. These data types are required to have been deposited in a public database in a specific format before the publication is accepted. Articles describing structures or sequences without a PDB identifier or UniProt/GenBank accession code will simply be rejected. If similar guidelines are adopted for mutations, an extremely valuable data set will become available. Concepts like RDF triplets [8] and nano-publications [9] are currently popular candidates for an ontology based publication method. Eventually, these guidelines will have to be extended to the publication of more complex data such as k_m and k_{cat} values for reactions, protein-protein interactions, etc. Systems like 3DM offer a glimpse of how much more efficient and productive the life sciences could become when data is generated, stored, and published using rigid standards. Future systems biology tools will benefit enormously from these structured knowledge bases, as much more data will be directly available instead of hidden away in articles.

Synthesis of Biochemicals

One of the main branches of protein engineering is the production of chemicals through the modification of existing enzymes. This is especially interesting for end products that are difficult, expensive or environment unfriendly to produce using non-biological processes.

The result of billions of years of evolution is a large number of proteins with a wide diversity in function. The ideal enzyme to produce a given chemical in an industrial setting is however seldom found. Even if such an enzyme is available other factors might prevent successful commercial application such as slow reaction speed, low affinity for the substrate, or host requirements unsuitable for large scale application. 3DM systems focus on the study of a super-family containing many related proteins with similar functionalities. Super-family data can be used by 3DM to quickly pinpoint functionally important amino acid positions for properties of interest. 3DM systems are therefore currently most productively used in protein engineering studies to improve the industrial usability of proteins. It is rewarding from a scientific point of view that 3DM systems have been shown to be capable of predicting amino acids positions that when mutated increase activity (chapter 3, [11]), tune substrate specificity (chapter 3, [12]), increase thermostability (chapter 7), and change enantio specificity (chapter 8, [13]).

DNA Diagnostics

In DNA diagnostics one tries to find the causal link between gene(s) and disorder. For monogenetic disorders this relationship can relatively easily be established using, for example, genome wide association studies. In a sense, the chance that a given mutation is causal for a genetic disorder is proportional to the product of three terms: the role of the residue in the protein, the role of the protein in the affected pathway, and the role of the pathway in the genetic disorder. In this thesis the focus has been mainly on the functional role of the amino acid position for the structure and for the function of the protein, and to a smaller extent on the role of the protein in its pathway. A non-synonymous 'mutation' in a disorder related gene is however not sufficient evidence to diagnose a patient for having the disorder. Differences between a sequenced gene from a patient and the 'reference' gene might simply be due to natural variation rather than a pathogenic mutation. Additionally, the phenotypes of many diseases overlap making it difficult to sort the 'pathogenic' from the 'benign' mutations. In chapters 4 and 6 for monogenetic disorders different methods are shown to distinguish natural variants from pathogenic mutations using the available data.

Many diseases however, are not directly related to a single gene but result from interactions between numerous proteins and metabolites in a complex system that operates on the same task or are part of the same pathway [14]. For polygenetic disorders, therefore, a systems biology approach is important to understand the biology of all aspects of the system, rather than the effects of mutations in the individual proteins. A full systems biology approach to study polygenetic diseases is not yet possible with current 3DM technology but combining 3DM systems for interacting proteins obviously would be the next logical step. A start to a more holistic approach has already been made. 3DM for example now can visualize data on external contacts such as protein-DNA or protein-ion contacts (chapter 2.3.11). In addition to these contacts, analysis of a mutation using HOPE (chapter 4) can detect whether a mutation might affect the regulatory domains of a protein. These heuristics are the

starting point for a systems biology approach to study polygenetic diseases on a higher level. Given enough time, we hope to be able to support such studies with 3DM like systems.

Drug Design

The third main field of interest for 3DM like systems is drug design. Drug design combines everything known about molecules to study and improve drugs but the process still relies very much on trial and error [4]. Several 3DM systems were made in collaboration with a pharmaceutical company (Organon now MSD) to study protein families of interest to pharma such as the Nuclear Receptor family. Unfortunately, research in pharma is seldom published, nevertheless the automatic data collection and integration in 3DM systems were able to significantly reduce the time required for certain projects [15].

Correlated Mutation Analyses

Generally literature provides only little information for a specific residue. In such cases one can check if any facts from literature are known about 'correlating' residues. Correlated residue positions co-evolve during evolution, that is when they mutate, they mutate together. In such cases one can assume that the residues are functionally linked and knowledge about the one residue can normally be directly transferred to the other residue. Trivially, if mutation of an arginine in a salt bridge causes a certain phenotypic effect, then the same effect will be expected on mutations of its partner. This is one of the reasons why Comulator (chapter 5) was developed. Examples of the usage of the Comulator tool are shown in chapters 3, 5, 7 and 8 where we show, for example, the use of correlated mutation analyses to improve substrate specificity and/or enzyme activity.

Outlook

One of the major tasks in the near future is to overcome some of the most important limitations we run into in both DNA diagnostics and protein engineering: the study of protein complexes. While working on the determination of the mode of interaction of Class B GPCRs with their endogenous ligands Florence Horn [16] showed the importance of synchronizing two MCSIS systems. The goal is thus to create 3DM systems for the individual parts of a pathway or protein factory, and organize the data in these 3DM systems so flexibly that we can interlink these systems and thereby facilitate studies such as the Class B GPCR ligand study. A chain of 3DM systems for proteins that are part of a pathway for example, might be used to determine if consecutively functioning enzymes need to physically interact or that each enzyme can just take the product of the previous step as its substrate. Inter-protein correlated mutation analyses could for example be used to highlight binding or protein interaction sites. In DNA diagnostics, such systems might be used for the study of polygenetic diseases to link multi-factorial causes to certain phenotypes. Eventually, we might even provide powerful inference engines for systems biology.

3DM was designed and built on the experiences gathered over many years in developing and using various MCSIS systems. Several publications have shown successful applications of 3DM systems in various enzyme engineering and DNA diagnostics projects. It is hoped that in the future 3DM systems can be used to easily and automatically extract, store and combine data and knowledge for complex pathways. With such powerful systems ever more complicated research projects can be facilitated.

References

1. S.M.G. Saerens, C.T. Duong, E. Nevoigt, Genetic improvement of brewer's yeast: current state, perspectives and limits, Appl. Microbiol. Biotechnol., 86 (2010) 1195–1212.
2. G. Festa, F. Autore, F. Fraternali, P. Giardina, G. Sannia, Development of new laccases by directed evolution: functional and computational analyses, Proteins, 72 (2008) 25–34.
3. T. Robak, E. Robak, New anti-CD20 monoclonal antibodies for the treatment of B-cell lymphoid malignancies, BioDrugs, 25 (2011) 13–25.
4. G. Seddon, V. Lounnas, R. McGuire, T. van den Bergh, R.P. Bywater, L. Oliveira, G. Vriend, Drug design for ever, from hype to hope, J. Comput. Aided Mol. Des., 26 (2012) 137–150.
5. R.K. Kuipers, H.-J. Joosten, R.H. Lekanne dit Deprez, M.M. Mannens, P.J. Schaap, Novel tools for extraction and validation of disease-related mutations applied to Fabry disease, Hum. Mutat, 31 (2010) 1026–1032.
6. R. Winnenburg, T. Wächter, C. Plake, A. Doms, M. Schroeder, Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies?, Brief. Bioinformatics, 9 (2008) 466–478.
7. A.J. Williams, L. Harland, P. Groth, S. Pettifer, C. Chichester, E.L. Willighagen, C.T. Evelo, N. Blomberg, G. Ecker, C. Goble, B. Mons, Open PHACTS: semantic interoperability for drug discovery, Drug Discovery Today, (2012).
8. X. Wang, R. Gorlitsky, J.S. Almeida, From XML to RDF: how semantic web technologies will change the design of "omic" standards, Nat. Biotechnol., 23 (2005) 1099–1103.
9. B. Mons, H. van Haagen, C. Chichester, P.-B. 't Hoen, J.T. den Dunnen, G. van Ommen, E. van Mulligen, B. Singh, R. Hooft, M. Roos, J. Hammond, B. Kiesel, B. Giardine, J. Velterop, P. Groth, E. Schultes, The value of data, Nat. Genet., 43 (2011) 281–283.
10. P.D. Stenson, E.V. Ball, K. Howells, A.D. Phillips, M. Mort, D.N. Cooper, The Human Gene Mutation Database: providing a comprehensive central mutation database for molecular diagnostics and personalized genomics, Hum. Genomics, 4 (2009) 69–72.
11. N.G.H. Leferink, M.W. Fraaije, H.-J. Joosten, P.J. Schaap, A. Mattevi, W.J.H. van Berkel, Identification of a gatekeeper residue that prevents dehydrogenases from acting as oxidases, J. Biol. Chem., 284 (2009) 4392–4397.
12. H.-J. Joosten, Y. Han, W. Niu, J. Vervoort, D. Dunaway-Mariano, P.J. Schaap, Identification of fungal oxaloacetate hydrolyase within the isocitrate lyase/PEP mutase enzyme superfamily using a sequence marker-based method, Proteins, 70 (2008) 157–166.
13. H. Jochens, U.T. Bornscheuer, Natural diversity to guide focused directed evolution, Chembiochem, 11 (2010) 1861–1866.
14. M.A. van Driel, H.G. Brunner, Bioinformatics methods for identifying candidate disease genes, Hum. Genomics, 2 (2006) 429–432.
15. S. Folkertsma, The nuclear receptor ligand-binding domain: from biological function to drug designa protein family-based approach, Radboud University Nijmegen, 2006.
16. F. Horn, R. Bywater, G. Krause, W. Kuipers, L. Oliveira, A.C. Paiva, C. Sander, G. Vriend, The interaction of class B G protein-coupled receptors with their hormones, Recept. Channels, 5 (1998) 305–314.

Summary

Summary

As the active component of many biological systems, proteins are of great interest to life scientists. Proteins are used in a large number of different applications such as the production of precursors and compounds, for bioremediation, as drug targets, to diagnose patients suffering from genetic disorders, etc. Many research projects have therefore focussed on the characterization of proteins and on improving the understanding of the functional and mechanistic properties of proteins. Studies have examined folding mechanisms, reaction mechanisms, stability under stress, effects of mutations, etc. All these research projects have resulted in an enormous amount of available data in lots of different formats that are difficult to retrieve, combine, and use efficiently.

The main topic of this thesis is the 3DM platform that was developed to generate Molecular Class Specific Information Systems (3DM systems) for protein superfamilies. These superfamily systems can be used to collect and interlink heterogeneous data sets based on structure based multiple sequence alignments. 3DM systems can be used to integrate protein, structure, mutation, reaction, conservation, correlation, contact, and many other types of data. Data is visualized using websites, directly in protein structures using YASARA, and in literature using Utopia Documents. 3DM systems contain a number of modules that can be used to analyze superfamily characteristics namely Comulator for correlated mutation analyses, Mutator for mutation retrieval, and Validator for mutant pathogenicity prediction. To be able to determine the characteristics of subsets of proteins and to be able to compare the characteristics of different subsets a powerful filtering mechanism is available. 3DM systems can be used as a central knowledge base for projects in protein engineering, DNA diagnostics, and drug design.

The scientific and technical background of the 3DM platform is described in the first two chapters. Chapter 1 describes the scientific background, starting with an overview of the foundations of the 3DM platform. Alignment methods and tools for both structure and sequence alignments, and the techniques used in the 3DM modules are described in detail. Alternative methods are also described with the advantages and disadvantages of the various strategies. Chapter 2 contains a technical description of the implementation of the 3DM platform and the 3DM modules. A schematic overview of the database used to store the data is provided together with a description of the various tables and the steps required to create new 3DM systems. The techniques used in the Comulator, Mutator and Validator modules of the 3DM platforms are discussed in more detail.

Chapter 3 contains a concise overview of the 3DM platform, its capabilities, and the results of protein engineering projects using 3DM systems. Thirteen 3DM systems were generated for superfamilies such as the PEPM/ICL and Nuclear Receptors. These systems are available online for further examination. Protein engineering studies aimed at optimizing substrate specificity, enzyme activity, or thermostability were designed targeting proteins from these superfamilies. Preliminary results of drug design and DNA diagnostics projects

are also included to highlight the diversity of projects 3DM systems can be applied to.

Project HOPE: a biomedical tool to predict the effect of a mutation on the structure of a protein is described in chapter 4. Project HOPE is developed at the Radboud University Nijmegen Medical Center under supervision of H. Venselaar. Project HOPE employs webservices to optimally reuse existing databases and computing facilities. After selection of a mutant in a protein, data is collected from various sources such as UniProt and PISA. A homology model is created to determine features such as contacts and side-chain accessibility directly in the structure. Using a decision tree, the available data is evaluated to predict the effects of the mutation on the protein.

Chapter 5 describes Comulator: the 3DM module for correlated mutation analyses. Two positions in an alignment correlate when they co-evolve, that is they mutate simultaneously or not at all. Comulator uses a statistical coupling algorithm to calculate correlated mutation analyses. Correlated mutations are visualized using heatmaps, or directly in protein structures using YASARA. Analyses of correlated mutations in various superfamilies showed that positions that correlate are often found in networks and that the positions in these networks often share a common function. Using these networks, mutants were predicted to increase the specificity or activity of proteins. Mutational studies confirmed that correlated mutation analyses are a valuable tool for rational design of proteins.

Mutator, the text mining tool used to incorporate mutations into 3DM systems is described in chapter 6. Mutator was designed to automatically retrieve mutations from literature and store these mutations in a 3DM system. A PubMed search using keywords from the 3DM system is used to preselect articles of interest. These articles are retrieved from the internet, converted to text, and parsed for mutations. Mutations are then grounded to proteins and stored in a 3DM database. Mutation retrieval was tested on the alpha-amylase superfamily as this superfamily contains the enzyme involved in Fabry's disease: an x linked lysosomal storage disease. Compared to existing mutant databases, such as the HGMD and SwissProt, Mutator retrieved 30% more mutations from literature. A major problem in DNA diagnostics is the differentiation between natural variants and pathogenic mutations.To distinguish between pathogenic mutations and natural variation in proteins the Validator modules was added to 3DM. Validator uses the data available in a 3DM system to predict the pathogenicity of a mutant using, for example, the residue conservation of the mutants alignment position, side-chain accessibility of the mutant in the structure, and the number of mutations found in literature for the alignment position. Mutator and Validator can be used to study mutants found in disorder related genes. Although these tools are not the definitive solution for DNA diagnostics they can hopefully be used to increase our understanding of the molecular basis of disorders.

Chapter 7 and 8 describe applied research projects using 3DM systems containg proteins of potential commercial interest. A 3DM system for the α/β-beta hydrolases superfamily is described in chapter 7. This superfamily consists of almost 20,000 proteins

with a diverse range of functions. Superfamily alignments were generated for the common beta-barrel fold shared by all superfamily members, and for five distinct subtypes within the superfamily. Due to the size and functional diversity of the superfamily, there is a lot of potential for industrial application of superfamily members. Chapter 8 describes a study focussing on a sucrose phosphorylase enzyme from the α-amylase superfamily. This enzyme can be potentially used in an industrial setting for the transfer of glucose to a wide variety of molecules. The aim of the study was to increase the stability of the protein at higher temperatures. A combination of rational design using a 3DM system, and in-depth study of the protein structure, led to a series of mutations that resulted in more then doubling the half-life of the protein at 60°C.

3DM systems have been successfully applied in a wide range of protein engineering and DNA diagnostics studies. Currently, 3DM systems are applied most successfully in project studying a single protein family or monogenetic disorder. In the future, we hope to be able to apply 3DM to more complex scenarios such as enzyme factories and polygenetic disorders by combining multiple 3DM systems for interacting proteins.

Samenvatting

De centrale plaats die eiwitten in veel biologische systemen innemen maakt hen van groot belang voor de levenswetenschappen. Eiwitten worden gebruikt in een groot aantal zeer diverse toepassingen zoals waterstof productie, opschonen van vervuilde grond, als medicijn, voor het diagnosticeren van patiënten, etc. Er is daarom veel onderzoek gedaan om eiwitten te karakteriseren en om de functionele en mechanistische eigenschappen beter te kunnen begrijpen. Onderzoekers hebben zich onder andere gericht op de opvouwing van eiwitten, reactie mechanismen, stress tolerantie, het effect van mutaties, etc. Al dit onderzoek heeft geresulteerd in een enorme hoeveelheid beschikbare data die aangeboden wordt door veel verschillende partijen en in veel verschillende formaten. Door deze enorme diversiteit is het lastig om de data automatisch te verzamelen, te combineren en te gebruiken in nieuwe projecten.

Hoofdonderwerp van dit proefschrift is het 3DM platform dat is ontwikkeld om automatisch databases voor eiwit families (Molecular Class Specific Information Systems, MCSIS/3DM systems) te kunnen genereren. Deze superfamilie systemen kunnen gebruikt worden om data te verzamelen en aan elkaar te koppelen middels op eiwit structuren gebaseerde sequentie-alignments. 3DM systemen bevatten onder andere eiwit, structuur, mutatie, reactie, conservering, correlatie, contact en vele andere soorten data. Al deze data wordt automatisch opgehaald, geconverteerd, aan elkaar gelinked en in een 3DM database opgeslagen. De data in een 3DM system wordt gevisualiseerd middels webpagina's, maar kan ook direct in eiwit structuren of artikelen getoond worden middels plugins voor YASARA en Utopia Documents. 3DM systemen bevatten een aantal tools zoals gecorreleerde mutatie analyses die gebruikt kunnen worden om de eigenschappen van de gehele familie of individuele positites en residuen te bestuderen. Om de eigenschappen van groepen eiwitten binnen de families te kunnen bestuderen en om groepen eiwitten met elkaar te kunnen vergelijken beschikt 3DM over de mogelijkheid om dynamische subsets aan te maken. Deze subsets bestaan uit een deel van de eiwitten in het systeem en kunnen op een aantal verschillende manieren gedefinieerd worden zoals op basis van de aanwezigheid van residuen op bepaalde posities, soort organisme, annotatie in structuur files, etc. 3DM systemen zijn daarom uitstekend geschikt om gebruikt te worden als centrale database voor protein engineering, DNA diagnostiek en drug design projecten.

De wetenschappelijke en technische achtergrond van het 3DM platform wordt beschreven in de eerste twee hoofdstukken. Hoofdstuk 1 beschrijft de wetenschappelijke achtergrond, beginnend met een overzicht van de basis principes van 3DM en daarna volgen beschrijvingen van hoofdmodules van het 3DM platform. De basis van elk 3DM systeem zijn de op structuren gebaseerde sequentie-alignments. Verschillende alignment methodes voor zowel structuur- als sequentie-alignments worden beschreven in combinatie met hun toepassingen en voor- en nadelen. Het 3DM platform bevat modules voor gecorreleerde mutatie analyses, mutatie extractie uit de wetenschappelijke literatuur en een pathogeniciteits

voorspeller voor onbekende varianten in genen die aan menselijke ziektes gekoppeld zijn. De achtergrond van gecorreleerde mutaties wordt behandeld aan de hand van verschillende implementaties. Problemen bij het automatisch verzamelen en opslaan van mutaties uit de literatuur worden behandeld. Hoofdstuk 2 beschrijft de architectuur van het 3DM platform vanuit een technisch perspectief. De 3DM database wordt beschreven aan de hand van een schematisch overzicht en een beschrijving van de inhoud van de belangrijkste tabellen. De benodigde stappen om een 3DM systeem op te zetten voor een superfamilie worden een voor een beschreven. 3DM modules voor gecorreleerde mutatie analyses (Comulator), mutatie extractie uit wetenschappelijke literatuur (Mutator) en het bepalen van de pathogeniciteit van mutanten (Validator) worden uitgebreid beschreven.

Hoofdstuk 3 bevat een beknopt overzicht van het 3DM platform, de verschillende modules binnen het platform en de resultaten van projecten gericht op het verbeteren van eiwitten. Dertien 3DM systemen zijn gegenereerd waaronder de PEPM/ICL en Nuclear Receptor superfamilies. Deze systemen zijn online beschikbaar en publiek toegankelijk voor vervolgonderzoek. Verschillende protein engineering studies zijn ondernomen aan de hand van deze 3DM systemen en zijn gericht op de optimalisatie van substraat specificiteit, het verhogen van enzym activiteit en het verbeteren van de stabiliteit bij hogere temperaturen.

Project HOPE, een tool voor het voorspellen van de effecten van een mutatie in een menselijk gen op de 3D structuur van het eiwit, wordt beschreven in hoofdstuk 4. Project HOPE wordt ontwikkeld aan de Radboud Universteit Nijmegen, Medical Center, onder supervisie van H. Venselaar. Een van de hoofddoelen van Project HOPE was het zoveel mogelijk hergebruiken van bestaande faciliteiten en databases middels het gebruik van webservices. Webservices maken het onder andere mogelijk om middels een gestandaardiseerde methode data op te halen en analyses elders uit te laten voeren. Dit vermindert de noodzaak voor krachtige lokale machines en databases. Na selectie van een mutant wordt data uit een aantal bronnen zoals UniProt en PISA verzameld. Een homology model wordt gegenereerd om de eigenschappen van residuen in de structuur te kunnen beoordelen. Effecten van de mutant op de structuur en op de functie van het eiwit worden beoordeeld met behulp van een beslissingsboom. Het resultaat is een overzichtelijk rapport waarin de effecten van de mutant worden beschreven.

Hoofdstuk 5 beschrijft Comulator: de 3DM module voor gecorreleerde mutatie analyse. Twee posities in een alignment correleren als deze ofwel allebei veranderen, oftewel geen van beide. Comulator berekent de correlatie tussen alignment posities binnen 3DM aan de hand van een statistical coupling algoritme. Statistical coupling is een techniek die frequentie van amino zuren op alignment posities analyseert en met elkaar vergelijkt om een gecorreleerde mutatie analyses score (CMA score) te berekenen. Correlaties worden in 3DM gevisualiseerd door middel van heatmaps en grafieken en kunnen in een eiwitstructuur getoond worden middels een plugin voor YASARA. Analyse van gecorreleerde mutaties in verschillende superfamilies toont aan dat sterk correlerende posities vaak in netwerken

voorkomen en dat de posities binnen deze netwerken vaak een gezamenlijke functie of eigenschap verzorgen. Aan de hand van deze karakteristieken zijn verschillende mutanten voorspeld om de specificiteit en activiteit van eiwitten te veranderen. Deze mutatiestudies bevestigen de voorspelde functies van de posities en tonen het nut van gecorreleerde mutaties aan voor rationeel eiwit design.

Mutator, de 3DM module voor text mining die gebruikt wordt om mutaties uit de wetenschappelijke literatuur te extraheren en in 3DM systemen op te slaan wordt besproken in hoofdstuk 6. Potentieel interessante artikelen worden geselecteerd via zoekqueries in PubMed. Deze artikelen worden gedownload, omgezet naar platte tekst en doorzocht op mutaties. De mutaties worden vervolgens gelinked aan eiwitten in een proces dat grounding wordt genoemd en opgeslagen in een 3DM database. Mutator is getest middels een 3DM systeem voor de alpha-amylase familie. Deze familie bevat het GLA eiwit waarvan bekend is dat mutaties leiden tot de ziekte van Fabry. In vergelijking met bestaande mutant databases zoals de HGMD en Swiss-Prot, bleek Mutator in staat om 30% meer mutaties uit de literatuur te verzamelen. Een veelvoorkomend probleem bij het voorspellen van de effecten van mutaties zijn de vele natuurlijke varianten die in een gen kunnen optreden zonder groot effect op de functie van het eiwit. Om pathogene mutanten van natuurlijke variatie te kunnen onderscheiden is daarom de Validator module aan 3DM toegevoegd. Validator gebruikt superfamilie data uit 3DM systemen om de pathogeniciteit van onbekende varianten te voorspellen. Onder andere conserveringsdata, toegankelijkheid van de zijketen in de structuur en het aantal mutanten dat voor een positie uit de literatuur is gehaald, kunnen gebruikt worden voor pathogeniciteits voorspellingen. Mutator en Validator bieden niet de definitieve oplossing voor het karakteriseren van onbekende varianten, maar zij kunnen hopelijk gebruikt worden om de moleculaire oorzaak van ziekte in beeld te brengen en de kennis over pathogene mutanten te vergroten.

De hoofdstukken 7 en 8 beschrijven 3DM systemen voor eiwitten met mogelijke commerciële en industriële toepassingen. Hoofdstuk 7 beschrijft het alpha-beta hydrolases 3DM systeem dat bestaat uit ruim 20.000 eiwitten met een grote diversiteit aan functies. Naast de structuur alignment voor de beta-barrel die alle eiwitten gemeenschappelijk hebben zijn er ook structuur alignments gemaakt voor 4 andere subtypes binnen de hydrolase familie. Elk van deze subgroepen heeft een unieke eigen conformatie naast de gezamenlijke beta-barrel. Het aantal eiwitten binnen deze superfamilie en de diversiteit aan functies van deze eiwitten resulteert in een groot aantal potentiële targets voor eiwit engineering projecten. Verschillende vervolgstudies die gebruik hebben gemaakt van het 3DM systeem hebben dit ook aangetoond. Tot slot beschrijft hoofdstuk 8 een project gericht op sucrose phosphorylase uit de alpha-amylase superfamilie. Dit eiwit kan potentieel op industriële schaal gebruikt worden om glucose moleculen te koppelen aan diverse andere moleculen. Doel van het onderzoek was het verbeteren van de stabiliteit van het eiwit bij temperaturen zoals deze vaak in industriele processen gebruikt worden. Middels een combinatie van rationeel eiwit design en structuur studies bleek het mogelijk om de halveringstijd van het eiwit op 60°C meer dan te verdubbelen.

3DM systemen zijn succesvol toegepast in een groot aantal verschillende protein engineering en DNA diagnostiek projecten. 3DM systemen zijn uitstekend geschikt gebleken voor gebruik in karakterisering en optimalisatie studies voor eiwit families en monogenetische ziektes. In de toekomst hopen we 3DM uit te breiden, zodat ook complexere systemen zoals enzyme factories en polygenetische ziektes middels geschakelde 3DM systemen onderzocht kunnen worden.

Acknowledgments

Na vijf jaar, vier werkplekken, meerdere werkgevers, diverse geldstromen en dankzij de hulp van een heleboel mensen is het gelukt! Ik heb in de afgelopen jaren met veel mensen samengewerkt, een hoop geleerd en een boel plezier gehad dus een paar bedankjes zijn hier wel aan de orde.

Als eerste mijn begeleiders vanuit Wageningen, Vitor en Peter. Vitor jij bent halverwege het traject mijn supervisor geworden en ik wil je hartelijk bedanken voor het begeleiden van deze promotie. Jouw systeembiologie invalshoek heeft een boel toegevoegd en onze discussies resulteerde altijd in veel nieuwe ideeën. Peter, in verschillende capaciteiten hebben wij al weer jaren met elkaar te maken. Eerst als stagebegeleider, daarna als hoofd van de Bio-Informatica opleiding, als adviseur van Bio-Prodict en de laatste paar jaar als begeleider van dit PhD traject. Jouw adviezen, correcties, ideeën en inzichten zijn van onschatbare waarde geweest in de ontwikkeling van 3DM en het succesvol afronden van dit proefschrift.

Gert, al vele jaren ben jij de drijvende kracht achter het MCSIS concept. Jouw drive, visie en enorme kennis van zaken hebben ervoor gezorgd dat de technologie is uitgegroeid van een concept tot een serie high-impact artikelen, meerdere PhDs en een succesvolle startup. Jou wil ik hartelijk bedanken voor de bemoedigende woorden als ik het even niet meer zag zitten, de eindeloze correcties, verbeteringen en notities op mijn schrijfwerk, je kennis en adviezen en voor alle gezelligheid op het CMBI en tijdens de barbecues in Malden.

Henk-Jan, wat jaren geleden bij jou begon als HBO stage heeft dan uiteindelijk in een PhD geresulteerd. Jij zocht destijds studenten om het maken van superfamilie systemen te automatiseren en daar zijn we na al die jaren nog steeds hard mee bezig. Hartelijk bedankt voor alle adviezen, de eindeloze discussies over soms zinnige en vaak onzinnige onderwerpen, diners, conferenties en natuurlijk voor het oprichten van Bio-Prodict! Wat klein begon is ondertussen tot een gezonde onderneming uitgegroeid waar we met z'n allen met recht trots op mogen zijn. De rest van het Bio-Prodict team, Tom en Bas en de studenten, Giorgio en Bram jullie bedankt voor de goede, gezellige en productieve werksfeer. Daarnaast wil ik iedereen waar ik de afgelopen jaren met veel plezier mee heb samengewerkt bedanken voor de hulp, tips, ideëen en discussies over van alles en nogwat: Jules, Maarten, Wilmar, Elmar, Tim, Hanka, Kal, & Coos bedankt!

Naast een PhD is het natuurlijk ook belangrijk om af en toe even te ontspannen met een potje kolonisten, of met een stuk varen op de Zorg. Erik, Ana, Nardy, Saïd, Rob, Bas, Cozmina, Mioumiou & Tim, bedankt voor alle gezellige middagen, etentjes, en weekendjes weg. Iedereen bij scouting in Nieuwegein ook van harte bedankt voor alle gezelligheid de afgelopen 20 jaar. Ik ben er afgelopen winter bijzonder weinig geweest, maar dat gaat komend jaar goedkomen! Thomas, Erik, Tom, Remy, Robbert, Niels, Rolo en al die anderen waar ik mee op kamp ben geweest, leiding van ben geweest, activiteiten mee heb gedaan en bier mee heb gedronken, bedankt!

Als laatste mijn lieve familie: Pap, Mam hartelijk dank voor al jullie goede zorgen, advies, steun en liefde in de afgelopen 30 jaar. Roos hartstikke bedankt voor al het werk wat je aan de layout en vormgeving van dit boekje besteed hebt, het is een mooi geheel geworden. Bien, Hayo, Mandy, Han, & Boef van harte bedankt voor alle gezellige etentjes, uitjes, verjaardagen en vakanties!

‘t is mooi
‘t is goed
‘t is gedaan
‘t is op
‘t is rond
‘t is klaar!

Curriculum Vitae

Remko Kasper Patrick Kuipers was born on August 13th, 1982 in Utrecht and grew up in Nieuwegein. He obtained a degree in office automation, from the lower grade school Scutos in Utrecht in 2002. Further specialising in computer science, he obtained his bachelor of Information and Communication Technology, with specialization in software development, at the Avans institute of professional education in 's Hertogenbosch in 2005. His internship was done at the Laboratory of Fungal Genomics, Wageningen University where he was involved in the initial design and development of what became the 3DM platform. He obtained a masters degree in Bio-Informatics at the University of Wageningen in 2007. For his MSc thesis Remko joined the CMBI group of prof. Vriend at the Radboud University in Nijmegen for a project focused on using webservices to assemble protein superfamilies. Remko has remained involved in 3DM since its inception, and has been employed by Bio-Prodict since 2009. From 2008 until 2012, Remko worked on his PhD in a cooperation between the Laboratory of Systems- and Synthetic Biology, Wageningen University, the Centre for Molecular and Biomolecular Informatics, Radboud University Nijmegen, and Bio-Prodict.

List of publications

Structure and function of 2,3-dimethylmalate lyase, a PEP mutase/isocitrate lyase superfamily member. Narayanan B, Niu W, Joosten HJ, Li Z, Kuipers RK, Schaap PJ, Dunaway-Mariano D, Herzberg O. J Mol Biol. 2009 Feb 20;386(2):486-503.

Correlated mutation analyses on super-family alignments reveal functionally important residues. Kuipers RK, Joosten HJ, Verwiel E, Paans S, Akerboom J, van der Oost J, Leferink NG, van Berkel WJ, Vriend G, Schaap PJ. Proteins. 2009 Aug 15;76(3):608-16.

3DM: systematic analysis of heterogeneous superfamily data to discover protein functionalities. Kuipers RK, Joosten HJ, van Berkel WJ, Leferink NG, Rooijen E, Ittmann E, van Zimmeren F, Jochens H, Bornscheuer U, Vriend G, dos Santos VA, Schaap PJ. Proteins. 2010 Jul;78(9):2101-13.

The alpha/beta-hydrolase fold 3DM database (ABHDB) as a tool for protein engineering. Kourist R, Jochens H, Bartsch S, Kuipers R, Padhi SK, Gall M, Böttcher D, Joosten HJ, Bornscheuer UT. Chembiochem. 2010 Aug 16;11(12):1635-43.

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

Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. Venselaar H, Te Beek TA, Kuipers RK, Hekkelman ML, Vriend G. BMC Bioinformatics. 2010 Nov 8;11:548.

Increasing the thermostability of sucrose phosphorylase by a combination of sequence- and structure-based mutagenesis. Cerdobbel A, De Winter K, Aerts D, Kuipers R, Joosten HJ, Soetaert W, Desmet T. Protein Eng Des Sel. 2011 Nov;24(11):829-34.

Overview of Completed Training Activities

Discipline Specific Activities

Benelux Bioinformatics Conference, Ghent, Belgium	2005
Benelux Bioinformatics Conference, Liège, Belgium	2009
Bio-Trans, Bern, Switzerland	2009
Workshop Bio TextMining, Ghent, Belgium	2010
European Human Genetics Conference, Gothenburg, Sweden	2010
European Conference on Computational Biology, Ghent, Belgium	2010
Capita Selecta in Complex Disease Analysis, Leuven, Belgium	2010
Applied Genomics of Industrial Fermentation, Wageningen	2010
NBIC Conference, Lunteren	2011
European Human Genetics Conference, Amsterdam	2011

General Courses

Scientific Writing, WUR Language Services	2009
Systems biology course: Statistics of ~omics data, WIE	2010

Optionals

Fungal Genomics biweekly group meetings, WUR	2008-2010
Preparation research proposal, WUR	2009-2010
Courses and presentation, CMBI Nijmegen	2010-2012
Microbiology biweekly PhD/Postdoc meetings, WUR	2010-2012

Colophon

Cover design and layout by roosgeeftvorm (http://www.roosgeeftvorm.nl)

Image credits:
Image 1.3 from http://www.cryst.bbk.ac.uk/pps97/assignments/projects/leluk/project.htm,
Image 1.7 copyright Bio-Prodict,
Image 1.8 from http://www.major.iric.ca/MC-Sym/faq.html,
Image 1.9 courtesy of Oliveira *et al.*
Image 2.4 courtesy of Kourist *et al.*
Image 1.2, 1.6, 1.10, 1.12 & 2.8 created using YASARA (http://www.yasara.org)
Image 2.1 & 2.2 created using Dia (http://live.gnome.org/Dia)

Printing by Pack & Parcel (http://www.pack-parcel.com)

The work described in this thesis was financially supported by VLAG, Radbound University Nijmegen, and Bio-Prodict.