

**3DM:**

***From Data to  
Medicine***

**Henk-Jan Joosten**

*Promotor:* **Prof. dr. J.A. van den Berg**  
Hoogleraar genomics  
Wageningen Universiteit

*Co-promotor 1:* **Dr. P.J. Schaap**  
Universitair docent, genomics  
Wageningen Universiteit

*Co-promotor 2:* **Dr. J.M. Vervoort**  
Universitair docent, biochemie  
Wageningen Universiteit

*Promotiecommissie:* **Prof. dr. S.C. de Vries**  
Wageningen Universiteit  
**Prof. dr. G. Vriend**  
Radboud Universiteit Nijmegen  
**Prof. Dr. J. van der Oost**  
Wageningen Universiteit  
**Dr. R. van Schaik**  
Organon, Oss

*Dit onderzoek is uitgevoerd binnen de onderzoeksschool VLAG (Voeding, Levensmiddelentechnologie, Agrotechnologie en Gezondheid)*

# 3DM:

## *From Data to Medicine*

### Proefschrift

ter verkrijging van de graad van doctor  
aan de Universiteit van Wageningen  
op gezag van de rector magnificus  
Prof. Dr. M.J. Kropff,  
volgens besluit van College van Decanen  
in het openbaar te verdedigen  
op 7 december 2007  
des voormiddag's om 11.00 in de aula  
door

Hendrik Johannes Joosten

geboren op 23 augustus 1975 te Amersfoort, Nederland

H.J. Joosten – 3DM: from Data to Medicine – 2007

Ph.D. thesis Wageningen University, Wageningen, The Netherlands (2007). 109 p.

**ISBN: 978-90-8504-965-0**



## Contents

<b>Preface</b>		<b>1</b>
<b>Chapter 1</b>	General introduction.	<b>3</b>
<b>Chapter 2</b>	3DM: A new generation of molecular-class-specific information systems applied to four protein super-families. (Joosten HJ, Folkertsma S, Zimmeren van F, Lutje Hulsik DJ, Kuipers R, Ittmann E, Roijen E, Vriend G, Schaap PJ.)	<b>21</b>
<b>Chapter 3</b>	Identification of fungal oxaloacetate hydrolyase within the isocitrate lyase/PEP mutase enzyme superfamily using a sequence marker based method. (Joosten HJ, Han Y, Niu W, Du J, Vervoort J, Dunaway-Mariano D, Schaap PJ.)	<b>41</b>
<b>Chapter 4</b>	Oxaloacetate hydrolase: The C-C bond lyase of oxalate secreting fungi. (Han Y, Joosten HJ, Niu W, Zhao Z, Mariano PS, McCalman MT, van Kan JA, Schaap PJ, Dunaway-Mariano D)	<b>61</b>
<b>Chapter 5</b>	Comulator: A new tool for calculating correlated mutations. (Joosten HJ, Akkerboom J, Kuipers R, Muthuraman M, Martens E, Oost van der J, Schaap PJ)	<b>83</b>
<b>Summary</b>		<b>101</b>
<b>Dankwoord</b>		<b>103</b>
<b>Curriculum Vitae</b>		<b>106</b>
<b>List of publication</b>		<b>107</b>
<b>Education</b>		<b>108</b>

## Preface

*Aspergillus niger* is famous for its ability to produce large amounts of organic acids. In the industry, for instance, *A. niger* is used for large scale citric acid production. Another organic acid that *A. niger* can secrete is oxalate. In fermentors, oxalate is a toxic unwanted by-product and it has been shown that oxalate secretion is an important factor in host infection by many (plant) pathogen fungi. The initial goal of the project described in this thesis was to increase our understanding of organic acid production. *A. niger* produces oxalate via the conversion of oxaloacetate into oxalate and acetate catalyzed by the enzyme oxaloacetate hydrolase (OAH). A compound that can inhibit OAH activity can potentially be used to prevent oxalate production in industrial processes and might be utilized as food protector by inhibiting oxalate production of plant pathogen fungi. It might even serve as a drug against the disease pulmonary oxalosis which is caused by fungi that infect the human lung cells and damage these cells by secreting large amounts of oxalate. A compound that can inhibit OAH can therefore be of significant economical value. However, the design of such an inhibitor is in general difficult.

As a student I worked at the CMBI (the bioinformatics department of the University of Nijmegen) on a project in collaboration with a Dutch pharmaceutical company. One of the main target proteins of this company is the nuclear receptor family (NR). As for many other protein families, various (internet) sources provide a wealth of freely available information. Pharmaceutical scientists that work on the development of drugs that target NRs, needed a bioinformatics system in which all available data for the NR protein superfamily is gathered and stored in a single database. Such databases allow for complex queries that can speed up the process of drug design. This approach, now known as “a superfamily based approach”, was published in the Journal of Molecular Biology (JMB). This article is the 4<sup>th</sup> hottest JMB article of July/September 2004 (<http://top25.sciencedirect.com/>).

At the start of this project not many details were known about OAH. Therefore, with my background, applying a superfamily based approach to OAH seemed a natural thing to do. However, to develop a superfamily system for OAH by hand is a huge effort and I decided to completely automate the generation of superfamily systems. The resulting program was called 3DM. Superfamily systems generated by 3DM proved to be powerful tools for the understanding and rational modification of proteins. Such systems can therefore potentially speed up drug design processes and guide rational protein design. Both topics will be addressed in this thesis. The OAH superfamily system was used as guidance to unravel protein-substrate interactions through directed mutagenesis. Understanding these interactions is important for drug design because drugs typically are small organic compounds that interact with a specific protein and thereby influence the activity of this protein. The protein engineering results obtained for OAH led to the design of a compound that strongly inhibits OAH.

This thesis describes all steps between the data collection and the final synthesis and *in vitro* testing of the OAH inhibitor. Therefore, this thesis is called 3DM: from Data to Medicine.

One of the tools developed in the course of this process is called Comulator, a statistical tool that can detect correlated mutations in superfamily alignments. To demonstrate the value of this tool various superfamilies were analyzed and the results were validated by mutagenesis experiments.



# Chapter 1

## General introduction

**Chapter 1****1. General Introduction***1.1 Background*

In the past century many proteins have been thoroughly studied, which resulted in useful insights into structure and function of those proteins and the usage of specific customized proteins in biomedical and biotechnological applications. In the pharmaceutical industry, for instance, proteins have become the main target for drug development. Drugs normally are small organic compounds that target specific binding sites in proteins that are involved in disease processes. On the other hand, the capacity of proteins, such as antibodies, to specifically bind to other proteins has led to their utilization in disease detection.

Enzymes are proteins that accelerate and facilitate chemical reactions. The ability of enzymes to catalyze (sometimes even stereo selective) very specific reactions make them appropriate tools for environmental friendly (bio)synthesis or breakdown of chemical compounds. Enzymes are therefore utilized in various industries, such as in the food processing industry where enzymes are traditionally used in the production of cheese, beer, wine, and baked products. In the last decade there is an increased usage of specific enzymes to improve the sensory properties of a product, involving taste, color, odor and touch and to replace harmful chemicals in food processing and preservation. Since the 1960s proteases and lipases are added to washing powder to remove stains that could otherwise only be removed by boiling and bleaching. In the paper industry enzymes are used to replace environmental unfriendly chlorine that is needed for the production of bleached paper. Phosphates that are present in animal waste are a major environmental problem in agriculture. Plants store phosphates as phytic acid: a single molecule with six phosphate groups. Poultry and pigs inefficiently release these phosphates from phytic acid. Phytases, that release phosphate groups from phytic acid, are added to animal feed to overcome this problem. Addition of phytases reduce the need for inorganic phosphorus supplementation, which lowers the level of harmful phosphates that are introduced to the environment through animal waste.

The environmental and economical advantages of industrial use of proteins are apparent, which justifies the exploration presented in this thesis: The development and application of a tool that can help scientist to improve (their knowledge on) proteins of interest. Such a tool can be useful for biotechnological and biomedical applications.

## Chapter 1

### *1.2 problem definition*

Protein engineering is often used to specifically modify physico-chemical parameters of proteins, such as solubility, stability or biochemical properties, such as ligand binding or substrate specificity. Principally there are two strategies used for protein engineering. The first is random mutagenesis. Here, mutations in a protein are randomly introduced followed by a screen for the desired effect. A review on the different random approaches was published in 2004<sup>1</sup>. Random mutagenesis does not require extensive knowledge about a target protein. A drawback of random mutagenesis is the huge number of mutations that have to be screened for a successful improvement. This makes this method laborious, expensive, time consuming and in some cases even impossible if high throughput screening is not possible. The second strategy is improvement through rational design, in which scientists try to predict mutations that lead to the changes in protein function. These predictions are normally derived from structural information of the protein. Due to the limited number of mutations normally needed for a successful improvement, rational design requires relatively inexpensive screening procedures. However, it can be extremely difficult to select target positions and to predict the effects of mutations at those positions based on structural information.

Many *in silico* strategies have been developed to improve the success rate of rational design<sup>2</sup>. Multiple sequence alignments (MSA), for instance, have been applied successfully in substrate binding sites prediction and the engineering of protein solubility. The use of MSAs for protein engineering will be discussed in section 2.

Proteins can be classified into protein superfamilies. The success rate of rational design of one specific member of a superfamily can be increased by the incorporation of bio-data from other members of this superfamily. At present, the cumulative amount of bio-data available for a specific superfamily is often large. However, this data is usually heterogeneous, often present in different formats and stored in multiple databases. Therefore, molecular class specific information systems (MCSIS) were developed<sup>3</sup>. Superfamily data gathered in a MCSIS is systematically stored in a unified format which makes this data transferable between protein members of interest. The MCSIS technology will be discussed in more detail in section 3.1. A number of popular MCSISes that are currently widely used, such as the NRMD<sup>4</sup> or GPCRDB<sup>4</sup>, are curated through human effort. From experience with these systems we know that human curated MCSISes are laborious to make and to keep updated. The 3DM program was developed to assist in this process as it automates the design, the creation and the updating of MCSISes. 3DM will be discussed in chapter 2.

## Chapter 1

### *1.3 Scope*

This thesis can be divided in two parts. The first is the development of 3DM. A general introduction on 3DM is given in section 3.3 and is discussed in more detail in chapters 2. 3DMs correlated mutation analyses method is discussed in chapter 5. The second part concerns the application of 3DM. Among five other superfamilies, the 3DM approach was applied to the enzyme oxaloacetate hydrolase (OAH) to predict effects of mutations in OAH (Chapter 3 and 4). An introduction on OAH and its superfamily is given in section 4.

### 2. Multiple sequence alignments

#### *2.1 MSAs and protein engineering*

For most proteins multiple homologs are present in public databases. MSAs built from these homologous protein sequences emphasize regions of similarity, which may indicate functional or evolutionary relationships between the proteins. Amino acid differences between correctly aligned homologous proteins can be interpreted as allowed substitutions. MSAs can therefore be seen as a sample of nature's successful mutagenesis experiments. The amino acid substitutions in MSAs demonstrate to which extent specific residues may be altered without destroying the structure and/or the function of the protein. Moreover, these substitutions provide insights into specific 3D positions that must be mutated in order to change the function of a protein. On the other hand, MSAs display indels (insertion/deletion events) that can be used to predict where structural variance is tolerated. MSAs therefore provide useful information that can guide rational protein engineering experiments. Another useful feature of MSAs is that they display co-evolution. To change the function of a protein usually multiple substitutions are needed. A random approach is unsuitable for the detection of a specific set of multiple mutations, since the introduction of multiple substitutions will exponentially expand the search space. In a MSA co-evolution appears as a network of alignment positions with simultaneous changing amino acid content. Positions that show co-evolution are normally functionally or structurally related, which can be used as guidance in protein engineering experiments. Co-evolution will be discussed in more detail in section 2.3 and examples of the use of co-evolution in the design of protein engineering experiments are given in chapter 3 and 5.

Another powerful method in which MSAs can be used for protein engineering is the transfer of information between proteins. Correctly aligned amino acids normally have a similar function within two aligned proteins and it is likely that they will show a similar effects if they are mutated. This characteristic of a MSA enables the transfer of amino acid related data, such as mutational information, from a well studied protein to a protein of interest. 3DM is developed for optimal use of all above described alignment characteristics.

#### *2.2 Structure based multiple sequence alignments*

Two or more homologues sequences can be aligned based on the primary sequence. Many different algorithms have been developed that can perform such sequence based alignments. The most popular one is the Clustal series of algorithms first published in 1988<sup>5</sup>. The quality of an alignment that is created by Clustal drops with a decreasing number of identical residues (sequence identity) that is shared by the sequences in the alignment. The performance of such algorithms decreases drastically if sequence identities drop below 30 % (Fig. 1, Fig. 2). Figure 1 shows an alignment performed by

ClustalW between two protein sequences with low sequence identity. The “true” alignment was derived from the superpositioning of the almost identical structures from these two proteins. The structural alignment shows that these sequences share only 44 identical residues (~20%). This example clearly shows that, in case of low sequence identities, it is much better to use structures for aligning proteins.

```

1DQU      MSYIEEDQRYWDEVAAVKNWUKDSRWRYTKRPFTAEQIVAKRGNLKIEYPSNVQAKKLW
1MUM      -----SLHSPGKAFRAALTKENPLQIVGTINANH-----

1DQU      GILERNFKNKEASFTYGCCLDPTMTVMQAKYLDTVYVSGWQSSSTASSTDEPSFDLADYPM
1MUM      -----ALLAQRAGYQAIYLSGGGVAAGSLG-----LPDLGI STL

1DQU      NTVPNHVNHLWMAQLFHDRKQREERMITPKDQRHKVANVDYLRPIIADADTGHGG-LTAV
1MUM      DDVLTHIR-----RITDVC SL-PLLVDADIGFGSSAFNV

1DQU      MKLTKLFVERGAAGIHIEDQAPGKKCGHMAGKVLVP ISEHINRLVAIRAQAD IMGTDLL
1MUM      ARTVKSMIKAGAA GLHIEDQVG-AKRCGHRPNKAIVSKEEMVDRI RAAVDARTDP--DFV

1DQU      ATARTDSEAATLITSTIDHRDHPFIIGSTNPDIQLNDLMVMAEQACKNGAELQAIEDEW
1MUM      IMARTDALAV-----

1DQU      EAPRTREGYYRYQGGTQCAINRAVAYAPFADLIWMESKLPDYKQAKEFADGVHAVWPEQK
1MUM      -----EGLDAAIERAQAYVEA GAEMLFPEAITE LANYRQFADAVQVPILANI

1DQU      LAYNLSPSFNWKKAMPREDEQETIYIKRLGALGYAWQFITLAGLHTTALISDTFAKAYAKQG
1MUM      TFGATPLFTTDELR-----SAHVAMALYPLSAFRAMNRAA

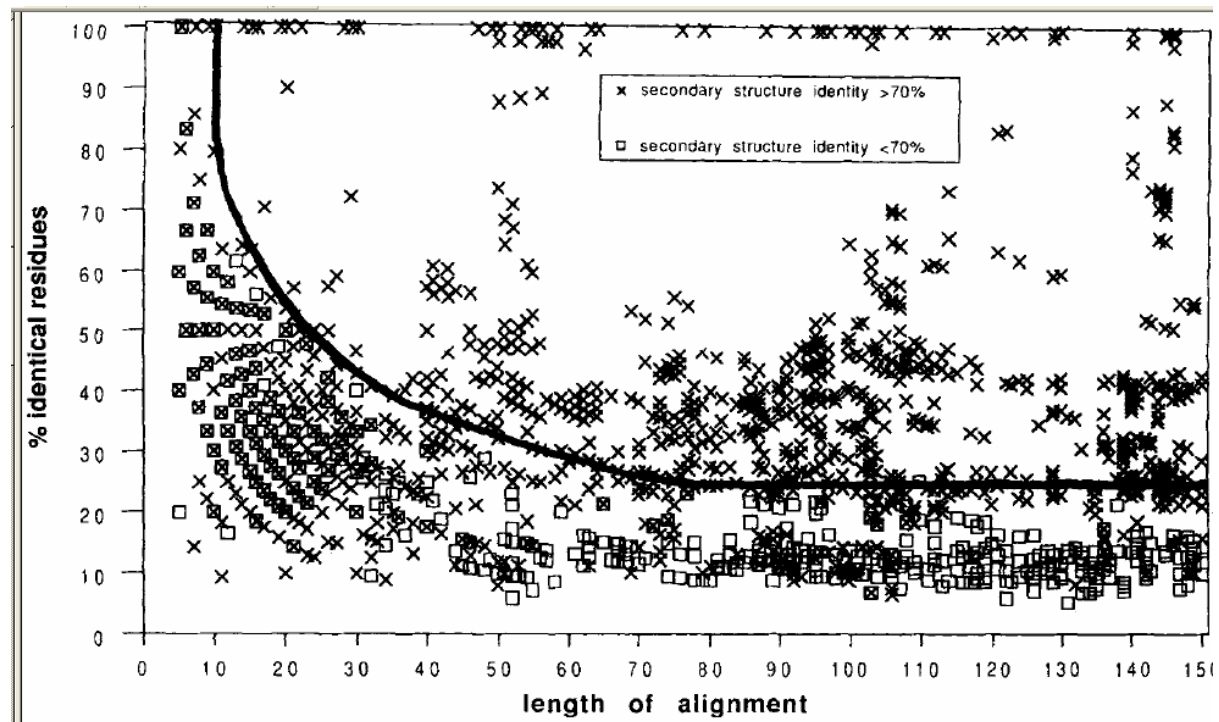
1DQU      MRAYGELVQPEMANGVDVVTHQKW SGANYVDNMLKMITGGVSSTAAMGKGVTEQFKS
1MUM      EHVVYNVLRQEGTQKSVIDTMQTR-----NELYESINYQYEEKLDNLFARSQVK-

```

**Fig. 1 Alignment of two sequences with very similar 3D structure (isocitrate lyase from *Aspergillus nidulans* (1DQU) and 2-methylisocitrate lyase from *Escherichia coli* (1MUM)), but with dissimilar sequences.**

The alignment was performed by ClustalW using default values. Only the Grey areas are aligned correctly.

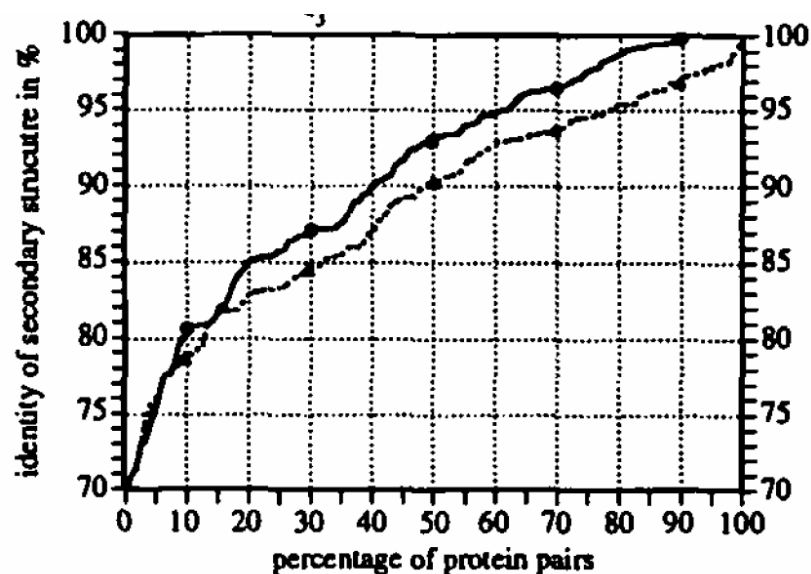
Many different algorithms, such as T-Coffee<sup>6</sup>, MUSCLE<sup>7</sup>, Kalign<sup>8</sup> have been developed to improve alignment performance. However, the creation of high quality alignments made from dissimilar (<30% sequence identity) sequences using the primary sequence only is very difficult or even an impossible problem to solve. This is illustrated in Fig. 2. Fig. 2 was published by Rost *et al.* in 1991<sup>9</sup> in a paper about homology modelling. For the generation of correct homology models accurate alignments are required. Models built from sequences of known structure can therefore be used to check alignment performance. From Fig. 2 it can be concluded that, based on sequence information alone, the sequences of two proteins of normal size (>150 residues) must share at least 30% identical residues to create a reliable alignment.



**Fig. 2 Threshold for reliable alignments as a function of alignment length.**

*This figure shows the minimal identical residues (y-axes) needed for acceptable alignments of different sequence length (x-axes) used for homology modelling. Each data point represents an alignment between 2 fragments of proteins of known structure. The black line represents an average of acceptable levels of identical residues. The upper right is the "save" zone in which the alignments resulted in correct homology models, whether the bottom left is the "danger zone" where too many alignments mistakes occurred for the generation of reliable homology models.*

Structurally conserved amino acids usually perform similar functions. A sequence alignment can be considered as an attempt to predict which amino acids occupy equivalent structural positions within the aligned proteins. Large accurate sequence alignments can therefore be used to understand the function of aligned amino acids because the function of an amino acid is dependent on the structural position it has within the protein. The structure of proteins is in general better conserved throughout evolution than the individual amino acid sequences. The sequence of a protein can change dramatically without changing the overall fold of the protein (Fig. 3). Proteins that are structurally 70% identical can have less than 5% identical amino acids. Therefore, if two protein sequences are similar, the structures of these proteins are normally highly similar. For aligning proteins, it would therefore be much better to use structural information instead of sequential information.



**Fig. 3 Relation between sequence conservation and structure conservation.** The structures of 126 non-homologues proteins<sup>9</sup> with similar 3D fold (70-100% identical(y-axes)) were used to determine the percentage of identical residues (x-axes) by superimposing these proteins. The two lines in the graph represent two different methods to determine structure conservation. From this graph it can be concluded that the structure of proteins is much better conserved than the amino acid sequence.

Figure from Rost *et al.* (1991)<sup>9</sup>

The power of a MSA increases with the quality of an alignment and the number of sequences in the alignment. To increase the number of sequences in a MSA the MSA can be built from more dissimilar sequences, but this usually results in an undesirable decrease of alignment quality. To bypass this problem MSAs can be built based on proteins of which the structure is solved. If enough structures are available a structure based MSAs (3DMSAs) can be built for a complete superfamily. The sequence identity between proteins in a superfamily can be as low as 5%. However, structurally these proteins can be conserved for more than 75%, which make these proteins “alignable” for 75% of their sequence (Fig. 3).

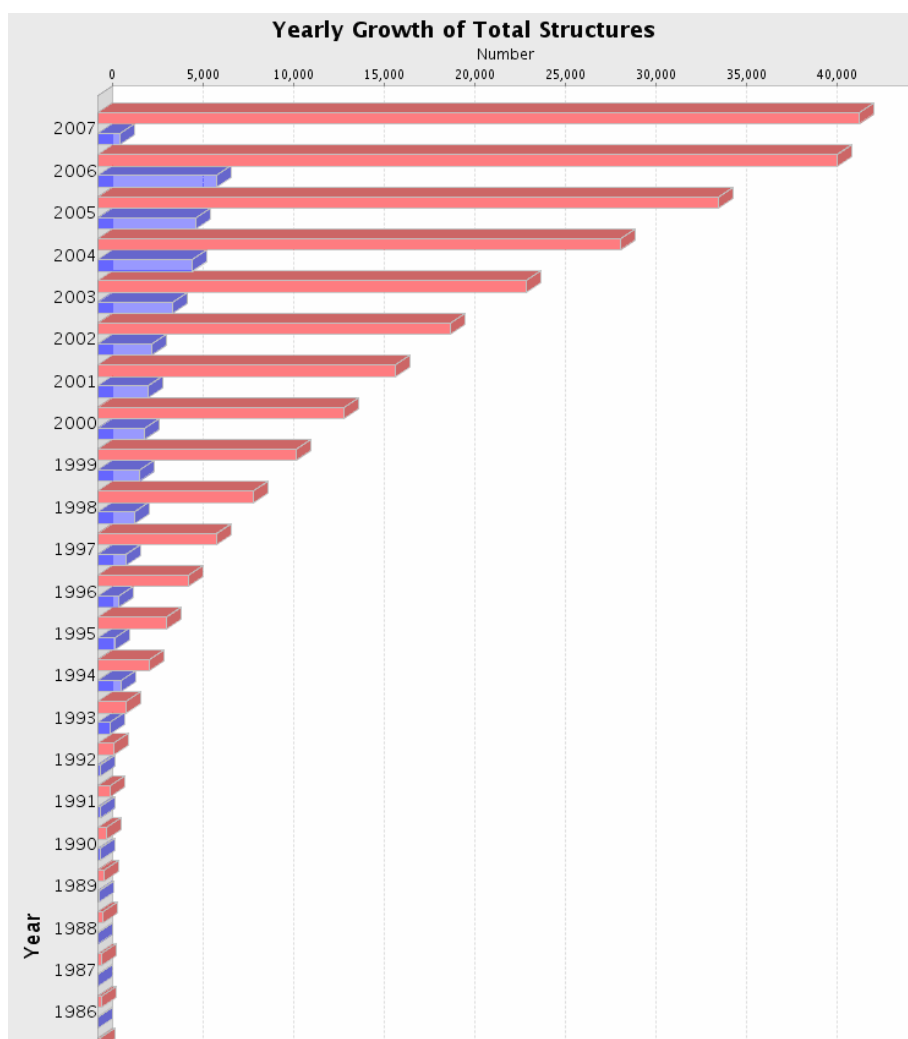
Due to a rapid increase in the speed of genome sequencing the number of sequences that can be collected has been and still is growing exponentially. In addition to the exponential growth in sequence information, large scale, high throughput, efforts in protein crystallization processes and NMR based structure determination have also resulted in an exponential growth of available high resolution protein structures (Fig. 4). Massive efforts have therefore been undertaken to classify protein structures based on their overall fold. SCOP (<http://scop.mrc-lmb.cam.ac.uk/scop/>), DALI (<http://www.ebi.ac.uk/dali/>) and CATH (<http://www.cathdb.info/latest/index.html>) are popular examples of protein structure classification databases. In these databases proteins with similar overall fold are classified into one superfamily. These classifications show that sequentially different proteins, which even may have different functions, can have equivalent three dimensional folds. The enzyme families described in this thesis, for example, contain enzymes that catalyze different reactions or even use different reaction mechanisms.



## Chapter 1

A superfamily 3DMSA, representing one specific fold, can contain more than 10,000 unique protein sequences and can be based on many (sometimes hundreds) protein structures. For protein engineering these large 3DMSAs are very powerful. They contain useful evolutionary fingerprints especially since the aligned sequences can have very dissimilar amino acid content and the functions of the proteins can be different.

In the first step of the creation of a 3DMSA protein structures belonging to the superfamily are superimposed. Residues that appear on top of each other in the superpositioning are also placed on top of each other in the alignment. In the second step, the sequences of the structures are used as templates to align very similar sequences (>40% identical) that can easily be aligned based on just the sequence. In this way separate subfamily alignments are created. These separate subfamilies can be aligned using the superpositioning of the first step, thereby creating a large superfamily alignment that can contain very dissimilar sequences. This two step procedure is discussed in more detail in chapter 2. This process can be very time consuming if done manually. 3DM is the first program that can automatically create superfamily 3DMSAs.



**Fig 4. Histogram representing the exponential growth of the total number of structures in the PDB database. Red bars indicate the total number of structures in the PDB database. The blue bars indicate the number of structures deposited to the PDB database in the corresponding year.**

### 2.3 Evolutionary fingerprints.

In this thesis it is shown that different types of information can be extracted from alignments and used for the design of site-directed mutagenesis experiments. The different types are conservation of amino acids at structurally conserved alignment positions, conservation of amino acids within specific subfamilies and correlated mutations. An introduction to these information types is given in the next paragraphs.

#### 2.3.1 Conservation of alignment positions

Proteins have functional constraints that limit the substitution rate of amino acids at each alignment position. Amino acid conservation is therefore directly linked to the contribution of the individual amino acids to protein function. Amino acids can have different functions within a protein. For example, a serine protease has a conserved catalytic triad that is responsible for hydrolysis of the peptide substrate. Mutating one of these residues of the catalytic triad will probably disrupt the catalytic activity. The same enzyme can have residues that are not directly involved in the reaction but are involved in positioning the substrate in the active site. Mutating these residues will probably only decrease the affinity of the enzyme for the native substrate. Other residues might be important for dimerisation or interaction with another protein in a supramolecular complex. Mutagenesis of the latter residues will probably have no effect on *in vitro* enzyme activity. Apart from giving structure to the protein, many residues in a protein do not have a clearly defined function and can almost freely be interchanged with another amino acid residue. Conservation of alignment positions can therefore be used to predict the function of specific amino acids within proteins and thus can be very helpful for protein engineering studies. In chapter 3 of this thesis alignment conservation applied to OAH is discussed.

#### 2.3.2 Subfamily specific residues

Protein superfamilies can be separated into functionally distinct protein (sub)families. Alignment positions holding conserved amino acid residues are believed to be involved in main functions common to all members. The proteins in each subfamily share specific properties that distinguishes them from other subfamilies, such as binding specificity of regulatory proteins, substrate specificity of enzymes, or transporter selectivity. The amino acids responsible for these subfamily specific functions are mostly conserved within a subfamily but can vary between the subfamilies. Mutating these residues might change specificity of the protein. Scanning alignments for subfamily specific residues can give guidance to protein engineering experiments. They can, for instance, help in the design of

## Chapter 1

novel enzymes that can be utilized in (environmental friendly) biochemical conversions of non-native organic compounds.

### 2.3.3 Correlated mutations

Most of the sequence analysis methods described in literature concentrate on alignment positions holding conserved amino acid residues and therefore focus on detection of residues involved in the main function of the protein. Besides the main function proteins contain other functional sites (see 2.3.1). In chapter 3 of this thesis the superfamily alignment of OAH was used to distinguish between alignment positions holding conserved residues and thus are important for catalysis and alignment positions holding a limited number of subfamily specific residues that contribute to substrate affinity. There are multiple ways to detect such positions. One way is screening the alignment for subfamily specific residues (section 2.3.2). To detect subfamily specific residues the alignment is screened for positions that are specifically conserved within a group of proteins that all have the same function. However, classification of proteins in subfamilies based only on their specific functions is often not possible, because the function many proteins is still not determined. An *ab initio* scoring method that does not require knowledge on protein function is to screen for alignment positions that show simultaneous amino acids substitutions. The rationale is that to execute different functions multiple adaptations in the protein sequence are required. In many cases these adaptations appear as group-wise substitution patterns (also known as correlated mutations) in the alignment. The next section introduces correlated mutation analyses as a tool to find such correlated mutations.

#### **CMA scoring example:**

In Fig. 5 five sequences are aligned (SeqA - SeqE) with residues at positions 1-4. Six possible position pairs can be formed from these four positions.

Position pair 1 and 2: These positions have a CMA high score, because between these two positions we exclusively observe simultaneous changes of amino acids (LM → FS → DR). The result is amino acid pairs having unique non-overlapping combinations.

Position pair 1 and 3: These positions also correlate, but the CMA score is not as high as between positions 1 and 2, because we observe overlapping combinations (The F in sequence E overlaps with the F in sequence C and D).

Position pair 1 and 4: The CMA score between positions 1 and 4 is equal to the CMA score between positions 1 and 3. The G in sequence E overlaps with the G in sequence A and B.

	1234		12	13	14	23	24	34
Seq A.	LMQG		LM	LQ	LG	MQ	MG	QG
Seq B.	LMQG		LM	LQ	LG	MQ	MG	QG
Seq C.	FSFM	pairs:	FS	FF	FM	SF	SM	FM
Seq D.	FSFM	→	FS	FF	FM	SF	SM	FM
Seq E.	DRFG		DR	DF	DG	RF	RG	FG

**Fig. 5 An alignment divided in all possible alignment position pairs.**

*A small alignment of 5 sequences (A -> E) with 4 residues (1 -> 4) (left side) is divided in all possible positions pairs (right side).*

All existing CMA methods compare 2 alignment positions for which a score for the significance of the correlation is calculated (CM score). The significance of a CM score not only depends on the quality of the alignment but also on the amino acid composition of the alignment. Methods, such as the mutual information method<sup>10,11</sup>, statistical coupling analysis<sup>12,13</sup> or the perturbation method<sup>14</sup>, compare the amino acid distributions to a random distribution and therefore require a normally distributed alignment. Alignments that consist of a large group of very similar sequences will give inequitable high CM scores for the alignment positions. Some algorithms, such as the Pearson correlation method<sup>15</sup>, contain certain alignment weight factors to bypass this problem, but mistakes in the alignment will always influence the outcome of these algorithms. To be able to screen for correlated mutations in the large 3DMSA's created by 3DM a new CMA algorithm was developed that could cope with alignment imperfections that are a result from the atomically creation of these alignments. This algorithm is discussed in chapter 5.

### 3. Superfamily databases

#### 3.1 3DM related databases

Nowadays different types of databases exist that can help scientists to design their protein engineering experiments. Many databases have been developed that store mutational information on protein(families). The Online Mendelian Inheritance in Man (OMIM; <http://www.ncbi.nlm.nih.gov/omim/>) database is a collection of mutations that are causing a wide spectrum of human diseases. More specialized databases have been designed that focus on mutational information for specific human proteins, such as the P53 tumor repressor protein. Three of these databases are discussed in chapter 2 of this thesis. Most of the existing mutational databases normally only contain mutational information that is specific for only one protein of specific species and not for a complete set of homologues proteins in a superfamily. The nuclear receptor mutation database (NRMD)<sup>16</sup> is an example of a mutational database designed for such a complete superfamily. In addition, most mutational databases contain only mutational data directly linked to a specific protein entry. Usually this data is not coupled to other types of (super)family data, such as alignment derived data or structural information.

On the other hand databases have been developed that store alignment related data. A good example is the HOMSTRAD database<sup>17</sup>. This database contains superfamily alignments made from over a 1000 superfamilies. These alignments provide information related to amino acids that can be extracted from the structure files, such as solvent accessibility. However, they are not linked to other types of amino acid related data, such as mutational information. The only systems that link all these different data types in one coherent system are the MCSIS type databases discussed in the next section.

#### 3.2 Molecular class specific information systems.

Using superfamily data to gain knowledge over a specific protein member, such as alignment derived data, structural information or mutational data, is known as a superfamily based approach<sup>18</sup>. The increasing amounts of available structural, sequential and biochemical data each flourishing in its own right, coming together in powerful new ways, make the superfamily approach a vital tool in rational protein design.

The drawback is that superfamily approaches are time consuming due to the heterogeneous nature of the available bio-data. The data is often in incompatible formats and scattered over multiple sources. In order to make useful correlations between these different data types and formats the superfamily data needs to be structured. Molecular class specific information systems (MCSISs) were designed for this purpose.

## Chapter 1

In the MCSIS technology all superfamily data are first collected from these heterogeneous sources and then stored in a MCSIS database together with alignment derived data and structural information. Analysis of the data occurs via a user- interface (mostly HTML pages). The G-coupled protein receptor database (GPCRDB) was the first MCSIS<sup>4</sup> with this setup. This database is now routinely used by scientists from all over the world and is visited over 8000 times/day. The second MCSIS database was designed for the nuclear receptors (NRDB)<sup>4</sup>. However, building these MCSISes requires much human intervention. Therefore, it is difficult to keep them up to date. To solve this problem 3DM was developed.

### 3.3 3DM

3DM is an automated analysis tool that can create MCSIS databases (3D-MCSISes). The data inside a 3D-MCSIS are specifically related to the amino acid sequences and information related to these sequences. Typical examples of data stored in a 3D-MCSIS database are protein sequences (and related data, such as the corresponding organism), protein structures and different types of amino acid related data, such as mutational information, phosphorylation sites, protein-protein interaction, substrate/ligand contacts, etc. To be able to transfer the amino acid related data between proteins of the superfamily, 3DM creates a structure based superfamily multiple sequence alignment and applies a so-called 3D-numbering scheme for all alignment positions. This 3D-numbering scheme is the core of a 3D-MCSIS. All amino acids and amino acid specific information are stored together with their 3D-number, which connects all the data enabling the transfer of data between proteins. Furthermore, the unique 3D-numbering scheme enables the scientist to make complex queries which enables easy correlation between different data types in a matter of seconds. The fully automatic nature of 3DM enables the easy creation of new 3D-MCSISes and allows for easy updating of an existing 3D-MCSIS. 3DM is discussed in chapter 2

### 4. OAH and its superfamily.

#### 4.1 Oxaloacetate hydrolase.

Oxaloacetate hydrolase (OAH) is an enzyme with a *B*-barrel fold that catalyses the conversion of oxaloacetate into oxalate (oxalic acid) and acetate. The involvement of OAH in oxalate production by fungi was already described in 1975<sup>19</sup>, but its gene was first cloned and characterized in 2000 from the fungus *A. niger*<sup>20</sup>. The second OAH that was cloned originated from the distantly related fungus *Ceriporiopsis subvermispota*<sup>21</sup> a fungus also capable of producing oxalic acid. Both groups that isolated the OAH encoding gene have not extensively characterized OAH. Substrate specificity was not fully characterized, no site-directed mutagenesis studies have been performed and the reaction mechanism of OAH was still unclear. Oxalate production was shown to be a virulent factor in host infection by fungi. An excellent review on this topic was published in 1996<sup>22</sup>. This thesis reports the correlation between the production of oxalate by fungi, the presence of an OAH encoding gene in their genome and the ability of fungi to be pathogenic. This correlation clearly shows the importance of oxalate production as a virulence factor. The design of a compound that can inhibit OAH can therefore be of great economical value.

#### 4.2 OAH-like protein class.

OAH belongs to a class of proteins, sharing high level of sequence similarity, ubiquitous present in filamentous fungi. The function of many of these proteins is still unknown, but in this thesis it is shown that many of these proteins are not OAH. Fungal genomes encode multiple OAH-like proteins. The recently sequenced genome of *A. niger*, for instance, encodes four OAH-like genes. Genetic evidence and sequencing of the *oah* gene of an oxalate non-producing strain of *A. niger* revealed that this gene is the only gene that encodes a true OAH. Due to the high sequence similarity it is helpful to have an OAH specific marker that enables to discriminate OAH encoding genes from closely related paralogs based solely on the primary sequence.

#### 4.3 Isocitratelase/PEP mutase superfamily.

OAH belongs to the Isocitratelase/PEP mutase enzyme superfamily. This superfamily contains enzymes catalyzing different reactions, some of which having totally different EC numbers, but they all catalyse a reaction in which a carbon-carbon bond is broken. Many members of the family have unknown substrates, but of six different subclasses the substrate is known. The first is isocitrate lyase (ICL, EC-number 4.1.3.1). This enzyme connects the citric acid cycle with the glyoxylate cycle by the reversible cleavage of isocitrate into glyoxylate and succinate. The reaction mechanism of this enzyme

## Chapter 1

was revealed in 1987<sup>23</sup> and the three dimensional structure was first resolved in 2000<sup>24</sup>. The second enzyme family is the closely related 2-methylisocitrate lyase (MICL, EC-number 4.2.1.99) that catalyzed the breakdown of 2-methylisocitrate. This protein was first crystallized in 2002<sup>25</sup>. A third characterized protein in this superfamily is the enzyme phosphoenolpyruvate mutase (PEPM, EC-number 5.4.2.9) catalyzing the isomerase reaction in which a phosphate group of phosphoenolpyruvate is rearranged resulting in 3-phosphonopyruvate. This enzyme was crystallized in 2002. A fourth protein is the very similar carboxyvinyl-carboxyphosphonate phosphorylmutase (CPEP, EC 2.7.8.23). The structure of this type of proteins is not available yet. Petal death protein, first crystallized in 2005<sup>26</sup> is a member of the OAH class of proteins. The available structures of proteins that belong to different subfamilies make this superfamily suitable for a 3DM approach.



### 5. Outline of this thesis

This thesis comprises of five chapters. **Chapter 1** is this general introduction. **Chapter 2** describes the method how 3DM was applied to 4 superfamilies. One of the superfamilies is the NR superfamily. In the year 2000, a MCSIS was built for the NR superfamily. **Chapter 2** shows a comparison of this system to the MCSIS build by 3DM three years later. This comparison clearly shows the importance of an up to date system. A 3D-MCSIS was made for the P53 superfamily. Using this superfamily we demonstrate the correlation between conservation of amino acids in an alignment and the intensity of the effect of introducing mutations in the human P53 protein. This study shows the power of such MCSISes for medical sciences. The third protein family for which a 3D-MCSIS was made is the Lama antibodies. This example shows how 3D-MCSISes can be used to understand structural features of proteins. The fourth MCSIS was made to understand OAH. **Chapter 2** describes how this 3D-MCSIS was used to find OAH specific amino acids. In this chapter, it is briefly discussed how this knowledge was used to clarify the unknown reaction mechanism of the enzyme and how this led to the design of an inhibitor. **Chapter 3** focuses on the use of the OAH 3D-MCSIS. The PEP superfamily is described together with the characterization of a new fungal specific protein class; the OAH-like protein class. It is described how these proteins relate to other members of the superfamily, which led to the discovery of OAH specific amino acids. One of these amino acids, an active site serine, proofed to be a reliable marker to discriminate OAH encoding genes from closely related paralogs.

**Chapter 4** describes the cloning, purification and characterization of OAH of *Botrytis cinerea*. This enzyme was used for protein engineering studies, which provided insight in the catalytic mechanism and led to the design of an inhibitor.

**Chapter 5** describes the newly developed correlated mutation analyses algorithm that was designed for the implementation in 3DM. This algorithm was used to detect OAH specific residues.

## 6. References

1. Williams GJ, Nelson AS, Berry A. Directed evolution of enzymes for biocatalysis and the life sciences. *Cellular & Molecular Life Sciences* 2004;61:3034-3046.
2. Gunsteren van WF. The role of computer simulation techniques in protein engineering. *Protein Eng.* 1988;2;5-13.
3. Horn F, Bettler E, Oliveira L, Campagne F, Cohen FE, Vriend G. GPCRDB information system for G protein-coupled receptors. *Nucleic Acids Res* 2003;31;294-297.
4. Horn F, Vriend G, Cohen FE. Collecting and harvesting biological data: the GPCRDB and nuclearRDB information systems. *Nucleic Acids Res* 2001;29:346-349.
5. Higgins DG, Sharp PM. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene*. 1988;73:237-44.
6. Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol.* 2000;302;205-17.
7. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 2004;32:1792-7.
8. Lassmann T, Sonnhammer EL. Kalign--an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics.* 2005;6;298.
9. Rost B, Sander C, Schneider R. Redefining the goals of protein secondary structure prediction. *J Mol Biol.* 1994;235;13-26.
10. Clarke ND. Covariation of residues in the homeodomain sequence family. *Protein Sci.* 1995;4;2269-2278.
11. Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW. Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. *Mol Biol Evol.* 2000;17;164-178.
12. Lockless SW, Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science.* 1999;286;295-259.
13. Fodor AA, Aldrich RW. On evolutionary conservation of thermodynamic coupling in proteins. *J Biol Chem.* 2004;279;19046-19050.
14. Dekker JP, Fodor A, Aldrich RW, Yellen G. A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments. *Bioinformatics.* 2004;20;1565-1572.
15. Gobel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. *Proteins.* 1994;18;309-317.
16. Van Durme JJ, Bettler E, Folkertsma S, Horn F, Vriend G. NRMD: Nuclear Receptor Mutation Database. *Nucleic Acids Res* 2003;31;331-333.
17. Mizuguchi K, Deane CM, Blundell TL, Overington JP. HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.* 1998;7;2469-2471.
18. Oxalate accumulation from citrate by *Aspergillus niger*. I. Biosynthesis of oxalate from its ultimate precursor. Muller HM. *Arch Microbiol.* 1975;103;185-189.
19. Pedersen H, Hjort C, Nielsen J. Cloning and characterization of oah, the gene encoding oxaloacetate hydrolase in *Aspergillus niger*. *Mol Gen Genet* 2000;263;281-286.
20. Folkertsma S, van Noort P, Van Durme J, Joosten HJ, Bettler E, Fleuren W, Oliveira L, Horn F, de Vlieg J, Vriend G. A family-based approach reveals the function of residues in the nuclear receptor ligand-binding domain. *J Mol Biol* 2004;341:321-335.
21. Yaver D, Cherry B, Murrell J. Polypeptides having oxaloacetate hydrolase activity and nucleic acids encoding same. 2004 Patent 6939701.
22. Dutton, Evans CS. Oxalate production by fungi: its role in pathogenicity and ecology in the soil environment. *Can J Microbiol* 1996;42:881-895.
23. Malhotra OP, Dwivedi UN, Singh J, Srivastava PK. Chemical reaction mechanism and active-site groups of isocitrate lyase. *Indian J Biochem Biophys* 1987;24;57-62.
24. Sharma V, Sharma S, Hoener zu Bentrup K, McKinney JD, Russell DG, Jacobs WR Jr, Sacchettini JC. Structure of isocitrate lyase, a persistence factor of *Mycobacterium tuberculosis*. *Nat Struct Biol* 2000;7;663-668.
25. Grimm C, Evers A, Brock M, Maerker C, Klebe G, Buckel W, Reuter K. Crystal structure of 2-methylisocitrate lyase (PrpB) from *Escherichia coli* and modelling of its ligand bound active centre. *J Mol Biol* 2003;328;609-621.
26. Teplyakov A, Liu S, Lu Z, Howard A, Dunaway-Mariano D, Herzberg O. Crystal structure of the petal death protein from carnation flower. *Biochemistry* 2005;44;16377-16384.

## Chapter 2

3DM: A new generation of molecular-class-specific  
information systems applied to four protein super-families

## **3DM: A new generation of molecular-class-specific information systems applied to four protein super-families**

### **1. Abstract**

Site-directed mutagenesis is often used to help increase our understanding of proteins. Many strategies have been developed to predict the function of amino acids and the effects of mutations. A multiple sequence alignment for a protein superfamily can be a powerful tool to transfer such information, but it also contains other relevant information about sequence variation and correlated mutations, for example.

3DM is a molecular-class-specific information system that creates an accurate structure-based multiple sequence alignment. Many derived data, such as correlated mutations, sequence variation, homology models, automatic mutation analyses, etc. are included. All of the information is stored in a relational database that revolves around a comprehensive 3D numbering scheme that encompasses all structurally equivalent positions, which allows the linking of all available data and the transfer of information between all sequences and structures.

The alignment is visualized and linked to derivative results via interactive HTML pages. The system provides a large series of computational protein engineering facilities, such as structure analysis and visualization, homology modeling, mutant structure and function prediction, etc. The possibilities of the system are illustrated using nuclear receptors, p53, phosphoenolpyruvate-mutase/isocitrate-lyase-like proteins, and single-chain antibodies as examples.

### 2. Introduction

In protein engineering and drug design, much can be gained when all available information about all members of a superfamily is available at the same time. For example, a mutation that influences antagonist binding in one nuclear hormone receptor (NR) is likely to have a similar effect in other NRs. Several WWW-interfaced information systems exist which harvest, store, validate, and present data for protein (super-)families. The GPCRDB<sup>1</sup> is a Molecular-Class-Specific Information System (MCSIS) for G protein-coupled receptors. This system, originally designed in 1993, has been a model for nuclear hormone receptors (NuclearDB)<sup>1</sup>, potassium channels (KchannelDB) and prions (PrionDB). These systems rely largely on manual curation, which makes it difficult to be up to date at all times. Nevertheless, the power of having so much heterogeneous data at one's disposal in one coherent system makes these systems popular.

Multiple sequence alignments (MSAs) are powerful tools in protein engineering because they can be used to predict the function of amino acids in one protein by transferring information from well studied members of a set of related proteins. The functional role of residues can also be revealed by analyzing the evolutionary information intrinsically present in an MSA, such as residue conservation, correlated mutations<sup>2</sup>, or entropy-variability patterns<sup>3</sup>.

Structural genomics projects rapidly populate the Protein Data Bank, and multiple structures will soon be available for nearly all protein families. 3DM was developed to exploit this and automatically creates the next generation of MCSISs. It generates an accurate structure-based MSA of a protein superfamily and builds homology models for all aligned proteins with unknown structure. It generates a 3D numbering scheme, which is used for all data, including crystal structures, sequences, and homology models. This enables collection, storage, hyper-linking, and easy correlation of many of these experimental and calculated data types. All data are mapped onto the alignment and can be retrieved via interactive HTML pages.

We have applied 3DM to a series of families: the NRs were included because this system allows comparison with the manually-curated NucleaRDB. 25.000 well-annotated p53 mutations are available, which makes this system a good test case for the automatic mutation analysis modules. The phosphoenolpyruvate-mutase/isocitrate-lyase-like (PEP/ICL) proteins were used because we can engineer these molecules in our laboratories, and thus validate the predictions. The variable domain of heavy chain antibodies<sup>4</sup> (VHH) were added because of their potential in biotechnological applications and their different structure function relations. Additionally, a large body of experimental data exists for VHHs. The largely automatic nature of 3DM allows us to update these 3D-MCSISs regularly.

### 3. Methods

#### 3.1 3DM

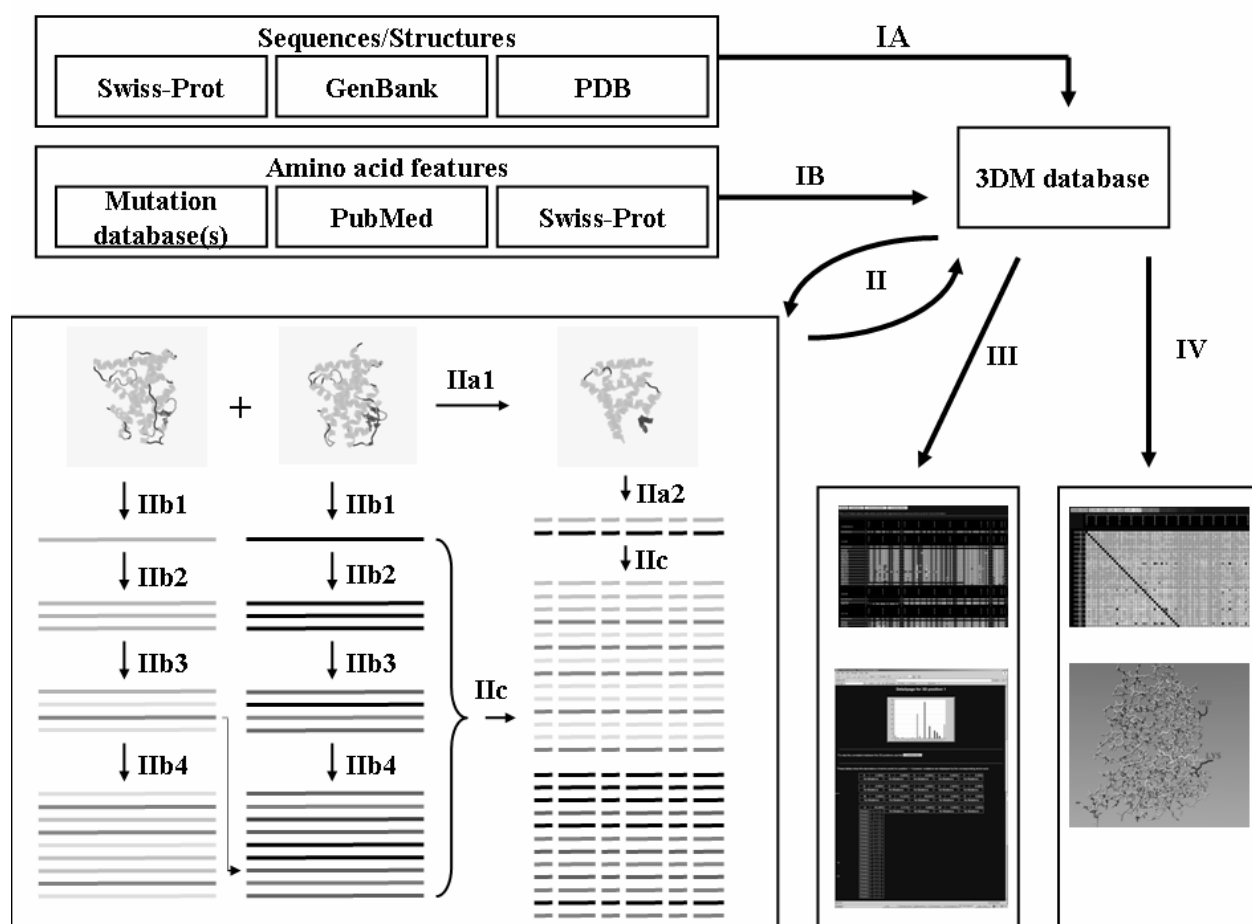
Figure 1 shows a schematic summary of the 3DM system. Database design, help files, documentation, etc., can be found at the 3DM homepage: <http://3dmcsis.systemsbiology.nl>. 3DM provides a series of user-adjustable parameters like sequence identity cut-off, percentage of structures that must participate 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 the five 3D-MCSISs discussed above. Optimization of these parameters is not yet possible for lack of CPU time.

#### 3.2 Data Collection

Data collection (Fig 1. Step I) 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/>), MuteXt<sup>5</sup>, three p53 mutation databases (<http://www-p53.iarc.fr/>, <http://p53.free.fr/>, [http://www.lf2.cuni.cz/projects/germline\\_mut\\_p53.htm/](http://www.lf2.cuni.cz/projects/germline_mut_p53.htm/)), etc.

Structural information was obtained by running BLAST<sup>6</sup> against the PDB, by extracting family members from SCOP<sup>7</sup>, and by using the EBI Secondary Structure Matching tool<sup>8</sup>. None of these systems is capable of extracting all relevant structures from the PDB, but when used in parallel, a nearly complete set can normally be obtained. BLAST searches against the NCBI nr database (<http://www.ncbi.nlm.nih.gov/BLAST/>) and the Swiss-Prot database (BLAST tool of ExPASy Proteomics Server, default values) were performed to obtain the sequences.

The NR sequences and structures were collected from the NucleaRDB. We selected the same 1577 sequences as used by Folkertsma *et al*<sup>9</sup> to allow comparison of the automatic generated system with the previously published, hand-curated system. The 1674 V<sub>H</sub>H sequences were obtained from the antibody variable domain database AVDDDB (containing both public and proprietary sequences; (<http://swift.cmbi.ru.nl/mcsis/systems/ABVDDDB/>)) together with the structural information from the PDB.



**Fig. 1 Summary of 3DM.**

Step I represents the collection of data. IA: Sequences and structures are automatically inserted into the 3DM database. IB: Features related to amino acids, such as mutational information, can be collected and inserted in the database automatically. Step II illustrates the creation of the 3D-MSA. Step IIb illustrates the iterative profile based alignment procedure 3DM uses to create the individual subfamily alignments. 3DM uses default five successive alignment rounds. In step IIc the separate subfamilies are aligned to one large superfamily alignment (for details see materials and methods). Note that sequences can be aligned to different structures here indicated by the arrow showing a sequence aligned in both subfamily alignments. This sequence is aligned to the grey starting sequence in step IIb3 and to the black sequence in step IIb4. This sequence therefore has a higher sequence identity with the grey sequence and is deleted from the black subfamily alignment in step IIc. Step III represents the visualization of the data in different HTML pages (See Fig. 3,4,5). Step IV illustrates the calculation and visualization of many alignment derived data such as correlated mutations (Fig. 2) and homology models (Fig. 7).

### 3.3 Multiple Sequence alignment

3DM uses a three-step procedure (Fig. 1, II) to create a superfamily alignment. In the first step (Fig. 1, step IIa1), all collected structures of a superfamily are superposed<sup>10</sup> and a common core of structurally equivalent positions is determined (Fig. 1, step IIa2). This 'core' is typically defined as the collection of residue positions where at least 90% of the structures have an alpha carbon located within 1.5 Ångström of the average position. In the second step (Fig. 1, IIb), separate subfamily alignments are created by aligning to each of the representative structures all sequences with more than typically 40% sequence identity. The alignments are created by WHAT IF<sup>11</sup> using the sequence-to-structure iterative

profile based alignment procedure as described by Oliveira *et al*<sup>12</sup>. Representative structures typically show less than 60% pair-wise sequence identity. Sequences that align well to multiple structures are used in the structure-based alignment only where they align best. In the third step (Fig. 1, IIc), these single-structure-based subfamily alignments are combined in one large alignment guided by the structure alignments from step one, i.e. residues that are structurally equivalent in the ‘core’ are forced to align. Each subfamily alignment comprises all residue positions present in its parent structure, whereas the ultimate combined alignment contains only ‘core’ residue positions.

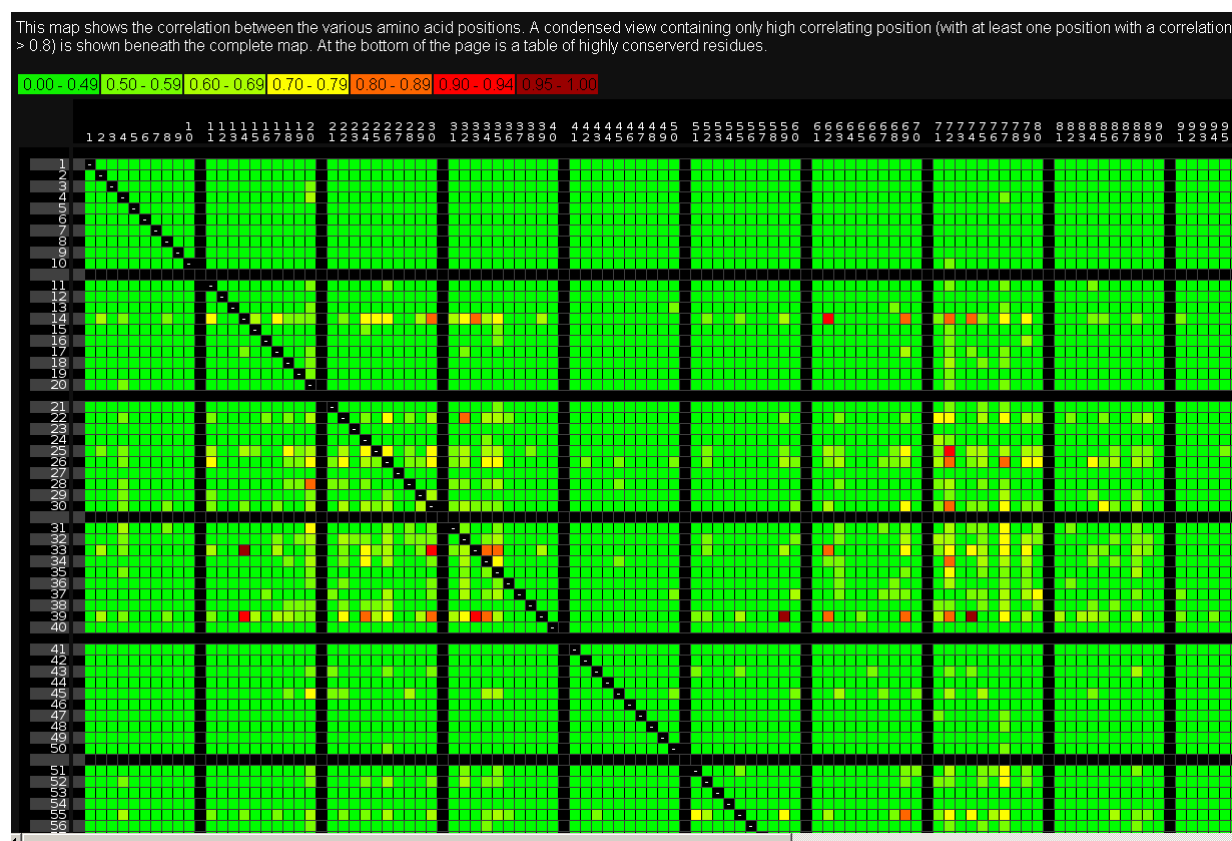
### 3.4 Numbering scheme

For all residues in the ‘core’, a numbering scheme is applied at the end of step II (Fig. 1). These residues are numbered sequentially starting at 1 for the most N-terminal structurally conserved residue position. The resulting numbers of the ‘core’ are called the 3D-numbers. These 3D-numbers are used throughout all alignments and are incorporated in all structure files and homology models. Sometimes, a commonly accepted numbering scheme exists for a superfamily. In the GPCRDB, for example, residues in helix I have numbers in the 100s, residues in helix II have numbers in the 200s, etc. 3DM can adopt such custom-made numbering schemes if desired.

### 3.5 Correlated mutations

Shulman *et al*<sup>16</sup> described an elegant correlated mutation analyses (CMA) method. 3DM determines CMA scores using a similar algorithm. The details of this algorithm are given in the 3DM homepage. The CMA scores of all positions are visualized in heatmaps (Fig. 2).





**Fig. 2** Screen shot of the correlated mutational analyses (CMA) heatmap of the NR superfamily. The numbers on the x- and y-axes are the 3D numbers. Positions with highest CMA scores are red.

### 3.6 Mutations/Features

Swiss-Prot holds many data about protein features such as post-translational modifications, mutations, natural variants, metal binding, etc. 3DM automatically collects and stores these features, and is sufficiently flexible to allow for the easy incorporation of data from many sources. For the NR superfamily, for example, additional mutation information could be collected using the MuteXt literature extraction system. For p53, the mutation information from three large collections: the IARC TP53 Mutation Database<sup>13</sup>, the UMD-p53 database<sup>14</sup>, and the Database of germline p53 mutations<sup>15</sup> were combined using small custom-made scripts.

### 3.7 Homology models

In step IV (Fig. 1) a homology model is built for each aligned protein with unknown structure, using the subfamily alignments from step IIb (Fig. 1) of the alignment procedure. Modelling is done with WHAT IF using the same default parameters as the server version<sup>17</sup>.

Search	Explanation	Amino Acid Detail	Correlation Map	Profiles										
Click on: Protein names, white amino acids in the alignment and consensus amino acids for more information														
CONSENSUS														
	1	0	2	2	3	4	5	6	7	7	7	8	8	
Consensus	A	L	L	L	S	L	L	A	P	T	F	P	A	
1A28A	1	0	2	2	3	4	5	6	7	7	7	8	8	
Consensus	L	L	T	S	L	T	S	L	L	A	P	T	F	P
1A28A	L	L	T	S	L	T	S	L	L	A	P	T	F	P
P06401	L	L	T	S	L	T	S	L	L	A	P	T	F	P
P06186	L	L	T	S	L	T	S	L	L	A	P	T	F	P
Q9GLW0	L	L	T	S	L	T	S	L	L	A	P	T	F	P
Q93449	L	L	T	S	L	T	S	L	L	A	P	T	F	P
Q00175	L	L	T	S	L	T	S	L	L	A	P	T	F	P
Q8W669	L	L	T	S	L	T	S	L	L	A	P	T	F	P
Q953K6	L	L	T	S	L	T	S	L	L	A	P	T	F	P
Q95390	L	L	T	S	L	T	S	L	L	A	P	T	F	P
Q28547	L	L	T	S	L	T	S	L	L	A	P	T	F	P
P07812	L	L	T	S	L	T	S	L	L	A	P	T	F	P
Q42274	L	L	T	S	L	T	S	L	L	A	P	T	F	P
Q91425	L	L	T	S	L	T	S	L	L	A	P	T	F	P
Q90009	L	L	T	S	L	T	S	L	L	A	P	T	F	P
Q91053	L	L	T	S	L	T	S	L	L	A	P	T	F	P
Q9UVY3	L	L	T	S	L	T	S	L	L	A	P	T	F	P
Q91172	L	L	T	S	L	T	S	L	L	A	P	T	F	P
Q91105	L	L	T	S	L	T	S	L	L	A	P	T	F	P
Q18391	L	L	T	S	L	T	S	L	L	A	P	T	F	P
Q93455	L	L	T	S	L	T	S	L	L	A	P	T	F	P
Q93879	L	L	T	S	L	T	S	L	L	A	P	T	F	P
Q902M7	L	L	T	S	L	T	S	L	L	A	P	T	F	P
Q91193	L	L	T	S	L	T	S	L	L	A	P	T	F	P
Q9V118	L	L	T	S	L	T	S	L	L	A	P	T	F	P
Q91AC6	L	L	T	S	L	T	S	L	L	A	P	T	F	P
P22199	L	L	T	S	L	T	S	L	L	A	P	T	F	P
Q91573	L	L	T	S	L	T	S	L	L	A	P	T	F	P
Q9N0W8	L	L	T	S	L	T	S	L	L	A	P	T	F	P
P08235	L	L	T	S	L	T	S	L	L	A	P	T	F	P
Q902M4	L	L	T	S	L	T	S	L	L	A	P	T	F	P
Q29131	L	L	T	S	L	T	S	L	L	A	P	T	F	P
Q90100	L	L	T	S	L	T	S	L	L	A	P	T	F	P
Q90HT2	L	L	T	S	L	T	S	L	L	A	P	T	F	P
Q9N0K3	L	L	T	S	L	T	S	L	L	A	P	T	F	P
Q18972	L	L	T	S	L	T	S	L	L	A	P	T	F	P
P79373	L	L	T	S	L	T	S	L	L	A	P	T	F	P
1BSXB														
	1	0	2	2	3	4	5	6	7	7	7	8	8	
Consensus	D	E	M	L	H	M	V	T	A	V	A	T	A	
1BSXB	D	E	M	L	H	M	V	T	A	V	A	T	A	
P10828	D	E	M	L	H	M	V	T	A	V	A	T	A	
P18113	D	E	M	L	H	M	V	T	A	V	A	T	A	
P37242	D	E	M	L	H	M	V	T	A	V	A	T	A	
P37246	D	E	M	L	H	M	V	T	A	V	A	T	A	
P37526	D	E	M	L	H	M	V	T	A	V	A	T	A	
P37243	D	E	M	L	H	M	V	T	A	V	A	T	A	
P18112	D	E	M	L	H	M	V	T	A	V	A	T	A	
Q96240	D	E	M	L	H	M	V	T	A	V	A	T	A	
Q02565	D	E	M	L	H	M	V	T	A	V	A	T	A	
P18117	D	E	M	L	H	M	V	T	A	V	A	T	A	
P18119	D	E	M	L	H	M	V	T	A	V	A	T	A	

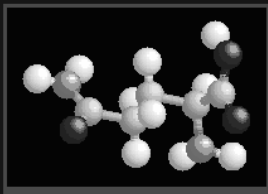
**Fig. 3** Screenshot of top part of the 3D-MSA of the ligand binding domain of the NR superfamily.

The two visible subfamilies are based on the structures 1A28, 1BSX. Residues are background coloured according to chemical property. For residues with a white background additional information is present in the 3DM database. These residues are directly linked to an amino acid detail page (Fig. 4). Residues in the consensus sequences on top of the alignment as well as on top of each subfamily alignment are linked to the position detail pages (Fig. 5). Finally, the protein identifiers on the left of the alignment are linked to protein detail pages.

### 3.8 Interface

Usage of a 3D-MCSIS generated by 3DM is organized via interactive WWW pages that interact with the underlying database (Fig. 1, step III). The homepage of 3DM is directly linked to the collection of alignments. An alignment (Fig. 3) is the main point of access to the data underlying each 3D-MCSIS, such as sequences, mutation data, 3D structures, etc., and to derived results such as CMA scores, 3D models, variability, and conservation in the alignment, etc.

**Glutamine 725 - alignment position 32**



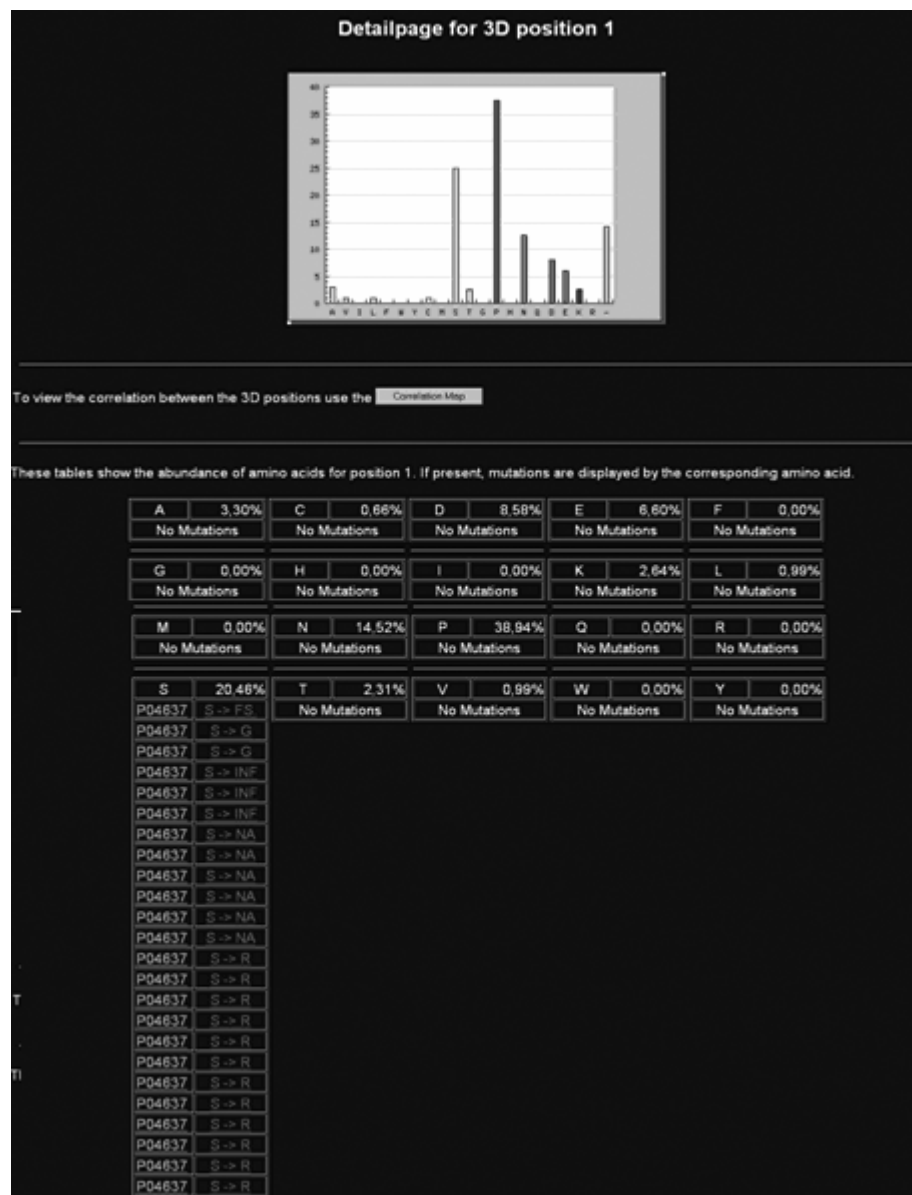
AminoAcid	Glutamine (GLN, Q)
Based on Class	1A28A (Round 1)
Protein	<a href="#">P06401</a>
Model	P06401
Location	Residue 12 of Core 2
Residue Number	725
3D Number	32
Profile Position	44
Build model with mutation on this position.	<a href="#">A</a> <a href="#">C</a> <a href="#">D</a> <a href="#">E</a> <a href="#">F</a> <a href="#">G</a> <a href="#">H</a> <a href="#">I</a> <a href="#">K</a> <a href="#">L</a> <a href="#">M</a> <a href="#">N</a> <a href="#">Q</a> <a href="#">P</a> <a href="#">R</a> <a href="#">S</a> <a href="#">T</a> <a href="#">V</a> <a href="#">W</a> <a href="#">Y</a>
<b>Features</b>	
MUTATION	Q > A
Description	
Link	<a href="#">PubMed</a>

[<< Previous Amino Acid](#)

**Fig. 4 Screenshot of an amino acid detail page.**

*Collection of all information of residue G725 of protein P06401 stored in the 3DM database.*

The alignments are linked to three different types of secondary page: protein detail pages, amino acid detail pages (Fig. 4), and position detail pages (Fig. 5). The protein detail pages provide information about the similarity of the sequences and their profile(s), the 3D models (superposed on the ‘core’ if required and numbered as desired), the raw sequence data, etc. 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. The position detail pages provide CMA results, amino acid distribution histograms, integrated positional mutation information from all sequences in the alignment, etc.



**Fig. 5 Screenshot of a position detail page of the p53 superfamily.**

The histogram shows the distribution of the amino acids at alignment position 1. The table at the bottom shows for position 1 the collection of mutational information stored in the 3DM database. For this position only mutational information was found for the human p53 (P04637).

### 3.9 Availability

The 3D-MCSISes are available from <http://3dmcsis.systemsbiology.nl/> free of charge.

### 4. Results and discussion

#### 4.1 Data collection

Information gathering (Fig 1. step I) is difficult, and the quality and completeness are hard to quantify. The collection of structure files, for example, is difficult because none of the available tools for obtaining similar 3D structures (e.g., SCOP, Structure Matcher, 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. As there is no definitive way to obtain a complete set of structure files, we opted to use many structure extraction methods in parallel to obtain as many PDB files as possible, even though this can result in some duplicated structures which must be removed later by the WHAT IF structure superposition software. Unfortunately, not all structures can always be used. Some structures contain too many ‘administrative’ errors to include them in the WHAT IF internal database used for superpositioning, and sometimes a unique superposition solution is hard to obtain because too many highly similar superposition solutions exist. The collection of sequences, on the other hand, is simpler because software like BLAST can easily be parameterized to include every possible homologue. In the MCSIS approach, correctness of alignments is much more important than completeness. Therefore, remote homologues sometimes cannot be used because the structures required for their reliable alignment are not yet available.

#### 4.2 Core

The determination of the core set of structurally superposed residue positions (Fig. 1 step IIa) is not necessarily free from differences in opinion or interpretation. The manually created NR alignment<sup>9</sup> resulted in a ‘core’ of 183 positions. 3DM automatically detected 158 of these residues. Most of the positions that were not detected by 3DM are involved in movement of helix I and helix XII in the NRs. These residues were added to the core by Folkertsma *et al.* even though they did not superpose well, but such decisions can, of course, not be made automatically. 3DM does allow the 3D-MCSIS curator to extend the alignments between steps IIa and IIb (Fig. 1).

#### 4.3 Sequence alignments

In 2004, Folkertsma *et al* collected 1577 NR sequences, 468 of which they could align reliably. Using this same set, 3DM can align 752 sequences. Although this increase is mainly a result of the larger number of structures available today, this increase demonstrates the importance of automation: the number of NRs in the multiple sequence alignment increased by 37% in two years without the need

for any extra human labour. This increase in number of alignable sequences results in a concomitant increase in usable information. For example, in 2004, no mutation information was available for 38 of the 158 common positions, whereas mutation information is now available for 30 of these 38. In nine cases, the gain of information is explained by the increased number of aligned sequences and in 21 cases by the increased availability of mutation information. In 2004, no mutation information was available for the conserved alanine of the WAK motif in helix III of the nuclear receptors, and contradicting predictions were made for the role of this residue. A recent mutation of this alanine (to a lysine) in the androgen receptor<sup>18</sup> was published and picked-up automatically. This mutation had an effect on ligand binding, but not on co-factor binding as high conservation of this alanine suggests<sup>9</sup>.

All 1647 V<sub>H</sub>H sequences can be aligned reliably using 3DM. The three gaps (between residues 25–26, 46–47, and 88–89) in the alignment correspond to the three complement determining regions (CDRs). V<sub>H</sub>Hs should be capable of recognizing an enormous number of antigens, so these CDRs display the highest variability. At least four main classes were detected using the alignment of the more conserved parts of these proteins (mainly the framework). A similar number of classes were described earlier in a study of 219 sequences with the CDRs included<sup>19</sup>.

### 4.4 p53

p53s are transcription factors involved in tumour repression. p53 mutations are found in more than 50% of all cancer patients<sup>20</sup> and in 580 different kinds of tumours<sup>15</sup>. The sequence of p53 is determined in many cancer patients, and more than 30.000 mutations detected this way have been collected and included in the p53 3D-MCSIS, making it the largest p53 mutation collection in the world. The phenotype (type of cancer) is known for almost all these mutations, allowing for completely new alleys of research. It is, for example, possible to relate residue conservation to the severity of disease after mutation. Fig. 5 shows that most amino acid types are allowed at ‘core’ position 1 in the 364 sequences of the p53 MSA. The twenty cases where a cancer is related to a mutation at this position all involve a new residue type not otherwise observed at this position. A thorough study of these effects is beyond the scope of this article, but this one simple example clearly illustrates the power of integrated heterogeneous data in the analysis of human disease. The observation that severity of phenotype is correlated with degree of conservation is very clear when comparing between core regions and structurally variable regions the number of point mutations that lead to cancer. The total number of mutations leading to cancer is almost 30.000. Of these, more than 25.000 are in ‘core’ positions. This is an average of 161 mutations per position, which becomes 382 when only positions more than 98% conserved are analysed. In comparison, in the structurally variable positions the average number of mutations per position that lead to cancer is only 22. Clearly, more conserved residues are more important, and are more likely to cause an adverse effect upon mutation. Table I shows the function of the 20 most conserved residues. Knowing that the zinc-binding region is important for DNA binding<sup>21</sup>,

## Chapter 2

it is safe to state that the mutation of conserved DNA binding residues is one of the surest paths to cancer.

Mutations	3D Number	conservation	Structural role
1725	116	98.6	DNA binding
1047	63	99.1	Near Zinc
718	150	98.7	DNA binding
486	67	100.0	His-Zinc bond
449	64	99.1	Cys-Zinc bond
404	46	99.4	
394	146	99.4	DNA surface
305	114	99.2	DNA binding
295	111	98.6	Cys-Zinc bond
278	149	98.5	DNA binding
240	145	98.5	DNA binding
239	21	98.6	
235	143	98.6	DNA binding
222	41	98.9	
172	27	98.6	Close to DNA surface
159	65	99.2	In zinc cluster
124	144	99.1	DNA binding
86	125	99.1	
67	31	99.4	
11	10	98.8	In DNA binding domain

**Table I: The frequency of mutations leading to cancer in the twenty most conserved alignment positions in p53 and their structural role.**

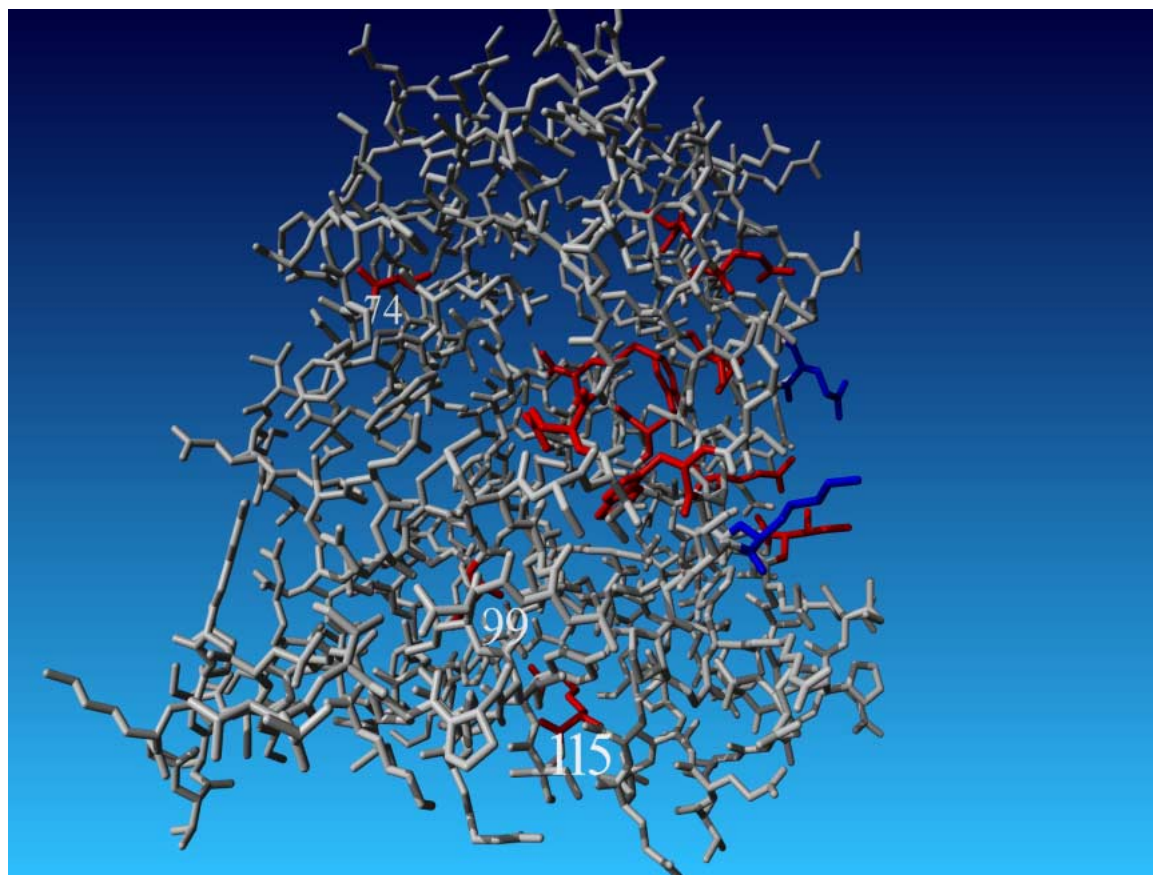
*The 3D number is the unique identifier given by 3DM to each alignment position. The structural role was inferred from visual inspection.*

### 4.5 NR

The amino acids of a nuclear receptor ligand-binding domain can be roughly divided into four groups<sup>9</sup>. Group 1 is involved in ligand binding (LB), group 2 in co-activator/repressor binding (CR), group 3 in dimerisation (DM), and group 4 contains amino acids that are involved in none of these inter molecular interaction related functions (NF). Mutational information is available for almost every position in the nuclear receptor superfamily. The 12 positions with a CMA score above 0.80 with at least two other positions (3D-numbers 14, 26, 30, 33, 39, 55, 59, 62, 66, 74, 99, 115) were analyzed. Only three of these positions (74, 99, and 115) are not located at the co-factor binding site (Fig. 6). This strongly suggests that the other nine positions are important for cofactor binding. Mutation information in the 3D-MCSIS confirmed this idea for seven of these nine residue positions. For the three positions not located at the cofactor binding site, mutational information is available only for position 115. Though not located at the cofactor binding site, mutation still leads to a complete loss of co-factor binding capability. No cofactor binding mutation information is available for positions 59, 62, 74, and 99. Position 62 was mutated (S. Folkertsma, manuscript in preparation) in an independent mutation screen for specific cofactor-binding residues. This mutation (Trp305Ala in the human Retinoid X receptor alpha) completely abolished cofactor binding. Position 59 is located very close to

the cofactor binding site and position 39, 59, and 99 are the three with the highest CMA scores in the nuclear receptor ligand binding domain. So, clearly, residues 59 and 99 are good candidates for future mutation studies.

The correlated mutation analyses produced only one probable false positive (position 74) out of twelve predicted functionally related positions. The fact that residue 14, in the difficult to align helix I, is correctly included in this set of residues with a high CMA score inspires extra confidence in the alignment procedure.



**Fig. 6 Representation of highly correlating positions in the NR ligand-binding domain.**

*Residues with high CMA scores are in red. Note that most of these residues (except for residues with 3D number 74, 99, 115) cluster together near the cofactor-binding site here marked by the conserved lysine of helix 3 (in blue) and the conserved glutamate of helix 12 (in blue). These residues form a charge clamp to which the cofactor can bind.*



### 4.6 $V_H Hs$

The relation between the functional roles of residues and their distribution of variability and conservation is different in  $V_H Hs$  from that in the other families studied. In the other four families, the active sites were characterized by high conservation while  $V_H Hs$  contain a series of highly variable positions, especially in their Complementarity Determining Regions (CDRs). In the 3DM alignment of  $V_H Hs$  thirteen positions are extremely well conserved. Cysteines at the positions 21 and 85 form the disulphide bridge that is conserved in  $V_H Hs$  as well as in the variable domains of conventional antibodies. A network of conserved core residues can be detected around this cysteine bridge. These residues have been experimentally shown to be important for the correct folding of the protein (D Lutje Hulsik, manuscript in preparation).

Significant CMA scores were found between the residue pairs 1:2, 4:10, 30:37, and 30:40. Positions 30, 37, and 40 correspond with the interface between the variable heavy chain (VH) and variable light chain (VL) in conventional antibodies. In conventional antibodies this interface consists of hydrophobic residues only. In  $V_H Hs$  these residues are more hydrophilic, as an obvious adaptation to the single chain status of  $V_H Hs$ <sup>22,23</sup>. This might suggest that  $V_H Hs$  evolved only recently from conventional antibodies, as discussed by Nguyen et al<sup>24</sup>.

The variability patterns clearly revealed that the boundary definition of these CDRs is similar to but not identical to the classical Chothia and Kabat definitions<sup>25,26</sup>. Mutational studies of residue 27 revealed that this position is part of CDR 1<sup>27</sup>. At the time of those studies, insufficient sequences were available to automatically delimit the CDR boundaries. Had the present ABVDDB-MCSIS been available in 2003, mutations at position 27 would not even have been attempted.

### 4.7 PEP/ICL

The PEP/ICL superfamily contains a variety of different enzymes that can break carbon-carbon bonds. The superfamily contains, for example, lyases (isocitrate lyase), hydrolases (oxaloacetate hydrolase), and mutases (phosphoenolpyruvate mutase). The enzymes catalyze widely varying types of reactions, but they all act on a bond between a carbon and an oxalate-like moiety. Several structures have been co-crystallized with an inhibitor, showing that the oxalate backbone of the substrate is bound to a metal ion ( $Mg^{2+}$  or  $Mn^{2+}$ ) in the active-site cleft (Fig. 7).

The positions 34, 41, 58, 62, 66, 68, 91, and 116 are conserved throughout all members of the PEP/ICL superfamily (Fig. 7). Five of these positions surround the metal ion and the oxalate moiety. The others are a proline (58) and two glycines (66, 68) and therefore seem more important for the dynamic aspects of catalysis than for the actual substrate binding. These eight residues are therefore likely to be very important for the reaction, and mutating any of them will lead to decreased activity.

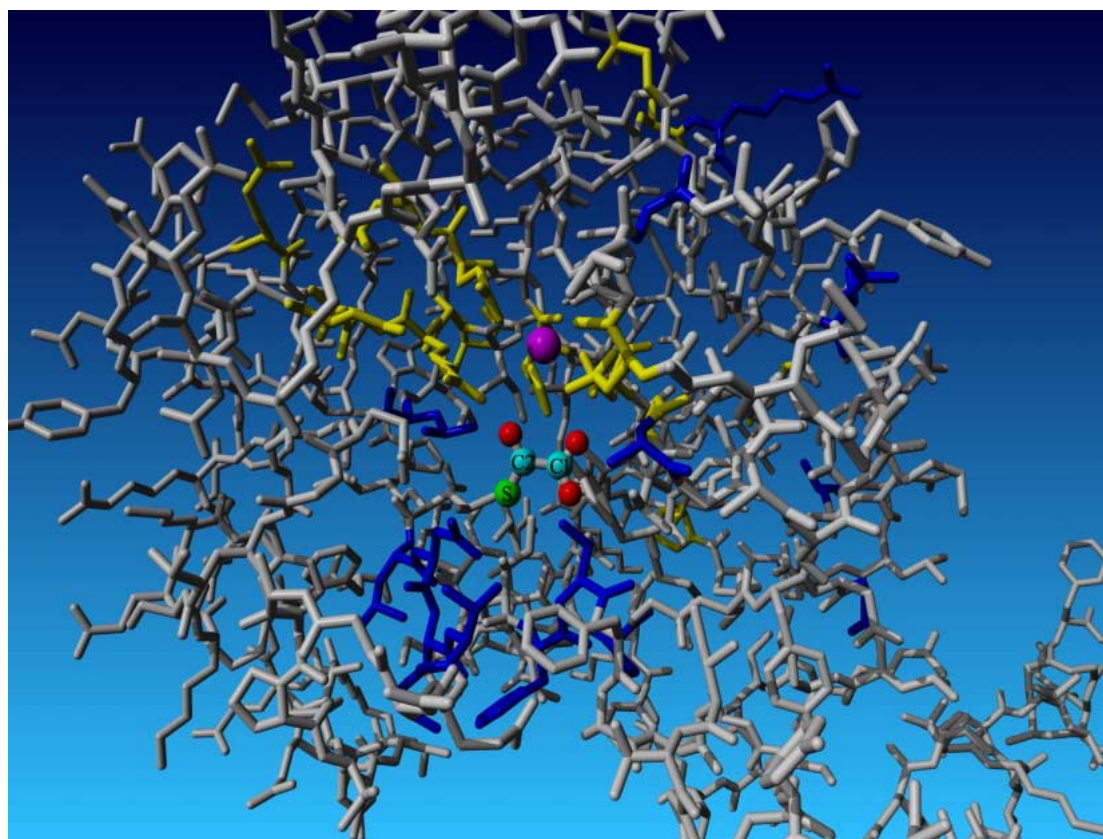
## Chapter 2

The mutational data section of the PEP/ICL 3D-MCSIS contains information for the positions 41, 62, 68, 91, and 116. Mutating any of these five positions strongly reduces enzyme activity.

Nine positions 43, 49, 75, 139, 147, 149, 150, 155, and 157 form a network of residues with high CMA scores. Most of these positions are located in the active-site cleft in the area where the substituent group of the substrate binds (Fig. 7). This suggests that this part of the active site is involved in substrate specificity. Since no mutational information could be found in the literature for any of these positions, we mutated one of them in the oxaloacetate hydrolase (OAH) of *Botrytis cinerea*. OAH converts oxaloacetate into oxalate and acetate.

We selected the position with 3D-number 157 (serine) since this position is closest in the homology model to the substituent group of the substrate. This S157 is a threonine or proline in most other sequences. We made three mutations: S157T, S157P, and S157A. These mutations did not significantly decrease  $V_{max}$ , but had drastic effects on the affinity of OAH for oxaloacetate (see chapter 3 of this thesis). Two more mutations were made in other proteins at this same position. One mutation was made in the petal death protein (PDP). PDP can convert multiple substrates including oxaloacetate. Like OAH, PDP<sup>28</sup> has a serine at 3D number 157. The same S157T and S157P mutations were performed, which caused the same effect in PDP (see chapter 3 of this thesis). Methyl isocitrate lyase (MICL) has a threonine at position 157. The complementary T157S mutation was made in MICL (data not shown). This mutation also had a drastic effect on the affinity of MICL for its substrate.

This shows how structural superfamily alignments can be used to transfer information about a well-studied protein to all other proteins in the alignment.



**Fig. 7 Conserved and highly correlated positions in the model structure of oxaloacetate hydrolase from *A. niger*.** Residues at conserved alignment positions (>97%) of the PEP/ICL superfamily are in yellow. These residues are located near the conserved oxalate moiety of the substrate and the conserved metal ion (purple). Residues at alignment positions having a high CMA score are in blue. Most of these residues form a network located near the substrate's substituent group (S) in green. The C1 carbon belongs to the acid group of the substrate's oxalate moiety that is conserved in the enzymes of the superfamily. The carbon of the conserved keto group is labelled C2.

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 highly variable.

This last example, which is worked out in detail in chapter 3 of this thesis, most beautifully shows the power of the use of data collected in a 3D-MCSIS. The multiple sequence alignment, the CMA, and the possibility of rapidly determining the correlation between residue type and enzyme class brought the previously undetected position 157 forcefully to our attention.

### **5. Conclusions**

The examples in this article show the power of integrated heterogeneous information about a superfamily stored in a database, such as a 3D-MCSIS database. The examples show that these databases can be used for the prediction of the function of amino acids in several ways. Amino acid functions can be revealed by using conservation, CMA results, or variability patterns. Amino-acid-related data such as mutational data can be transferred from a studied member of the superfamily to a protein of interest. The creation of an MCSIS involves interactions with many web-services, databases, and software packages, and can easily take several hours of CPU time. 3DM was therefore designed to create a 3D-MCSIS database as automatically as possible. The number of solved protein structures is increasing exponentially, so 3DM will rapidly become applicable for the study of an increasing number of protein families.

## 6. References

1. Horn F, Vriend G, Cohen FE. Collecting and harvesting biological data: the GPCRDB and nuclearDB information systems. *Nucleic Acids Res* 2001;29:346–349.
2. Oliveira L, Paiva AC, Vriend G. Correlated mutation analyses on very large sequence families. *Chembiochem* 2002;3:1010–1017.
3. Oliveira L, Paiva PB, Paiva AC, Vriend G. Identification of functionally conserved residues with the use of entropy-variability plots. *Proteins* 2003;52:544–552.
4. Hamers-Casterman C, Atarhouch T, Muyldermans S, Robinson G, Hamers C, Songa EB, Bendahman N, Hamers R. Naturally occurring antibodies devoid of light chains. *Nature* 1993;363:446–448.
5. Horn F, Lau AL, Cohen FE. Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors. *Bioinformatics* 2004;20:557–568.
6. 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.
7. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 1995;247:536–540.
8. Krissinel E, Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr D Biol Crystallogr* 2004;60:2256–2268.
9. Folkertsma S, van Noort P, Van Durme J, Joosten HJ, Bettler E, Fleuren W, Oliveira L, Horn F, de Vlieg J, Vriend G. A family-based approach reveals the function of residues in the nuclear receptor ligand-binding domain. *J Mol Biol* 2004;341:321–335.
10. Vriend G, Sander C. Detection of common three-dimensional substructures in proteins. *Proteins* 1991;11:52–58.
11. Vriend G. WHAT IF: a molecular modeling and drug design program. *J Mol Graph* 1990;8:52–56.
12. Oliveira L, Paiva ACM, Vriend G. A model for G-protein coupled receptors. *J Comp Aided Mol Des* 1993;7:649–658.
13. Olivier M, Eeles R, Hollstein M, Khan MA, Harris CC, Hainaut P. The IARC TP53 database: new online mutation analysis and recommendations to users. *Hum Mutat* 2002;19:607–614.
14. Beroud C, Soussi T. The UMD-p53 database: new mutations and analysis tools. *Hum Mutat* 2003;21:176–181.
15. Sedlacek Z, Kodet R, Poustka A, Goetz P. A database of germline p53 mutations in cancer-prone families. *Nucleic Acids Res* 1998;26:214–215.
16. Shulman AI, Larson C, Mangelsdorf DJ, Ranganathan R. Structural determinants of allosteric ligand activation in RXR heterodimers. *Cell* 2004;116:417–429.
17. Rodriguez R, Chinae G, Lopez N, Pons T, Vriend G. Homology modelling, model and software evaluation: three related resources. *Bioinformatics* 1998;14:523–528.
18. Alen P, Claessens F, Schoenmakers E, Swinnen JV, Verhoeven G, Rombauts W, Peeters B. Interaction of the putative androgen receptor-specific coactivator ARA70/ELE1alpha with multiple steroid receptors and identification of an internally deleted ELE1beta isoform. *Mol Endocrinol* 1999;13:117–128.
19. Harmsen MM, Ruuls RC, Nijman IJ, Niewold TA, Frenken LG, de Geus B. Llama heavy-chain V regions consist of at least four distinct subfamilies revealing novel sequence features. *Mol Immunol* 2000;37:579–590.
20. Glover-Kerkvliet J. p53 in 3-D. *Environ Health Perspect* 1994;102:1034–1036.
21. Celli J, Duijf P, Hamel BC, Bamshad M, Kramer B, Smits AP, Newbury-Ecob R, Hennekam RC, Van Buggenhout G, van Haeringen A, Woods CG, van Essen AJ, de Waal R, Vriend G, Haber DA, Yang A, McKeon F, Brunner HG, van Bokhoven H. Heterozygous germline mutations in the p53 homolog p63 are the cause of EEC syndrome. *Cell* 1999;99:143–153.
22. van der Linden RH, Frenken LG, de Geus B, Harmsen MM, Ruuls RC, Stok W, de Ron L, Wilson S, Davis P, Verrips CT. Comparison of physical chemical properties of llama VHH antibody fragments and mouse monoclonal antibodies. *Biochim Biophys Acta* 1999;1431:37–46.
23. Muyldermans S. Single domain camel antibodies: current status. *J Biotechnol* 2001;74:277–302.
24. Nguyen VK, Su C, Muyldermans S, van der Loo W. Heavy-chain antibodies in Camelidae; a case of evolutionary innovation. *Immunogenetics* 2002;54:39–47.
25. Chothia C, Lesk AM, Tramontano A, Levitt M, Smith-Gill SJ, Air G, Sheriff S, Padlan EA, Davies D, Tulip WR, et al. Conformations of immunoglobulin hypervariable regions. *Nature* 1989;342:877–883.

## Chapter 2

26. Kabat, E. A., Wu, T. T., Perry, H. M., Gottesmann, K. S. & Foeller, C. Sequences of Proteins of Immunological Interest. 1991;5th edit., NIH Publication no. 91-3242 U.S. Department of Health and Human Services
27. Dolk E., manuscript in preparation.
28. Teplyakov A, Liu S, Lu Z, Howard A, Dunaway-Mariano D, Herzberg O. Crystal structure of the petal death protein from carnation flower. *Biochemistry* 2005;44;16377-16384.

## Chapter 3

Identification of fungal oxaloacetate hydrolyase within the isocitrate lyase/PEP mutase enzyme superfamily using a sequence marker based method

## Identification of fungal oxaloacetate hydrolyase within the isocitrate lyase/PEP mutase enzyme superfamily using a sequence marker based method

### 1. Abstract

*Aspergillus niger* produces oxalic acid through the hydrolysis of oxaloacetate, catalyzed by the cytoplasmic enzyme oxaloacetate acetylhydrolase (OAH). The *A. niger* genome encodes four additional open reading frames with strong sequence similarity to OAH yet only the *oahA* gene encodes OAH activity. OAH and OAH-like proteins form a filamentous fungi subclass of the isocitrate lyase/PEP mutase enzyme superfamily. Analysis of function-specific residues using a superfamily based approach guided by a newly developed system called 3DM (see companion paper)<sup>1</sup> revealed an active site serine as a possible sequence marker for OAH activity. We propose that presence of this serine in family members correlates with presence of OAH activity whereas its absence correlates with absence of OAH. This hypothesis was tested by carrying out a serine mutagenesis study with the OAH from the fungal oxalic acid producer *Botrytis cinerea* and the OAH active plant petal death protein as test systems.

### 2. Introduction

Filamentous fungi, such as the food biotechnology fungus *Aspergillus niger*, the opportunistic human pathogen *A. fumigatus*, the phytopathogenic fungi *Botrytis cinerea* and *Sclerotinia sclerotiorum*, and numerous brown-rot and white-rot basidiomycetes are able to efficiently produce and secrete large quantities of oxalate<sup>1</sup>. Because oxalate is toxic (a concern in using fungi for commercial food and drug production) and a key factor in fungal pathogenesis<sup>2-5</sup>, it is important to distinguish oxalate-producing strains from non-producing strains. Oxalate can be formed from oxaloacetate in a C-C bond lysis reaction catalyzed by oxaloacetate hydrolase (oxaloacetate acetylhydrolase, OAH, EC 3.7.1.1)<sup>6</sup> and from the oxidation of glyoxylate<sup>7</sup> and glycolaldehyde<sup>8</sup>. The acetoacetate/OAH route predominates in oxalate producers examined to date<sup>1,9-14</sup> and therefore the *oah* gene might be used for the classification of fungi as potential oxalate producers.

Recently, the genomes of a large number of oxalate and non-oxalate producing fungi have been sequenced. The object of this work was to correlate the ability of fungi to produce oxalate with the presence of an *oah* gene in their genome. This task was made difficult by the fact that fungal genomes encode several OAH homologs having an unusually high level of shared sequence identity. Thus,



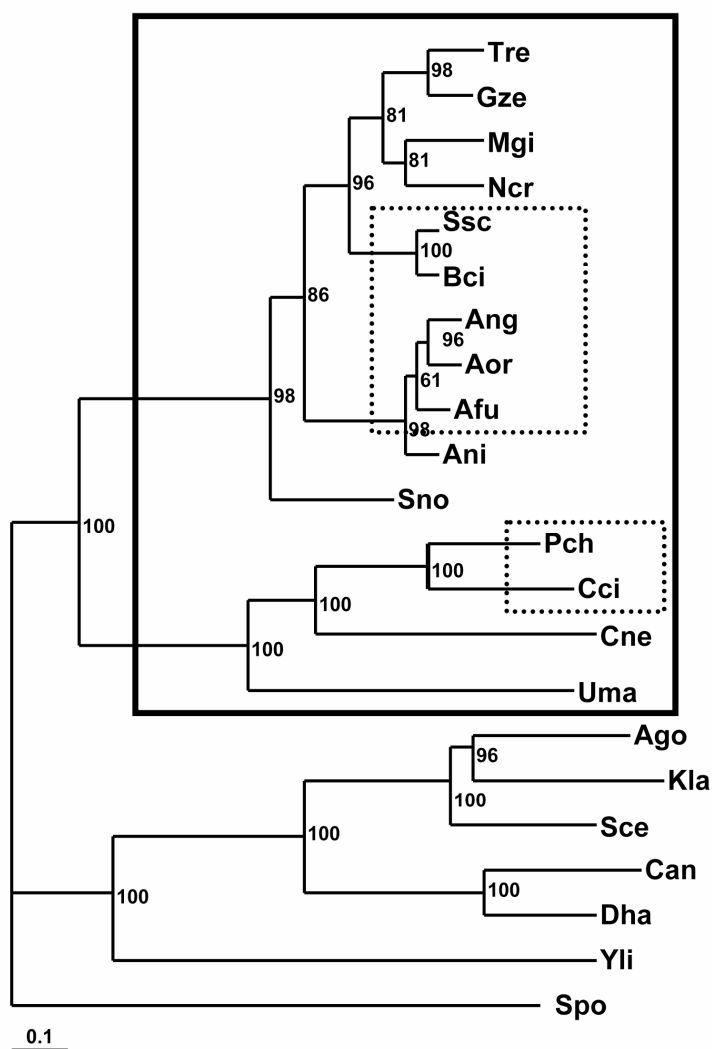
identification of OAH activity in the gene product would require purification of each protein followed by *in vitro* OAH activity testing. A less time consuming approach, based on structure-function analysis of OAH was therefore pursued. In this paper we propose the identification of a sequence marker for OAH activity that can be used to identify the *oah* gene in fungal genomes.

Family based approaches that combine 3D structure data with sequence analyses techniques are powerful methods to reveal the function of residues<sup>15</sup>. To identify OAH specific sequence markers this approach was applied to isocitrate lyase/PEP mutase superfamily of which OAH is a member. A large structure based superfamily multiple sequence alignment (3D-MSA) was built to which we have applied a general numbering scheme (3D-numbers) for structurally conserved positions. This 3D-numbering scheme will be used throughout this paper. From this structural alignment we were able to segregate sequences into subclasses, one of which we named the ‘OAH-like’ class of fungal proteins. The sequences of the members of this subclass from fungi with confirmed OAH activity<sup>12,16,17,18,19</sup> were shown to be distinguished by the presence of a serine, which from homology modeling and comparison to the recently reported structure of the OAH active plant petal death protein, we know to be located within the active site. By carrying out amino acid replacement experiments with recombinant OAH from *Botrytis cinerea* and with the recombinant OAH-active petal death protein from *Dianthus caryophyllus* (carnation)<sup>20</sup>, we tested the contribution that the active site serine makes in catalysis of OAH cleavage. Here we report the results from these studies which support the role of the active site serine in OAH catalysis, and which provide insight into the divergence of OAH activity within the  $\alpha$ -hydroxyacid C-C lyase branch of the isocitrate lyase/PEP mutase enzyme superfamily. Moreover, we show that the active site serine identified by analyses of the 3D-MSA is a reliable sequence marker for OAH activity.

### 3. Results and discussion

#### 3.1 Identification of the Fungal OAH and OAH-like subclass of the Isocitrate Lyase/PEP mutase Enzyme Superfamily.

The *Aspergillus niger* OAH protein sequence was used in BLAST searches for the identification of similar sequences within the *A. niger* predicted proteome, in the non redundant (NR) protein database, and in predicted proteomes of completely sequenced fungal genomes. The results are summarized in Table I. Most of the queried filamentous fungal genomes have one or more genes which appear to encode OAH-like proteins and have strong sequence similarity to the *A. niger oahA* gene (Table I) (Fig. 1).



**Fig. 1 Maximum likelihood phylogenetic tree of 22 completely sequenced fungi.** Fungal genomes encoding OAH-like class proteins are enclosed by a rectangle with a solid border. Rectangles with dotted borders enclose genomes encoding a (putative) OAH. For species abbreviations see Table I.

Species	Source db *	Abbr.	Protein Acc. No.	Taxonomy
Ascomycota (complete sequenced genomes with predicted proteomes)				
<i>Saccharomyces cerevisiae</i>	1	Sce	-	Saccharomycotina
<i>Kluyveromyces lactis</i>	1	Kla	-	
<i>Ashbya gossypii</i>	1	Ago	-	
<i>Debaryomyces hansenii</i>	1	Dha	-	
<i>Candida albicans</i>	1	Can	-	
<i>Yarrowia lipolytica</i>	1	Yli	-	
<i>Schizosaccharomyces pombe</i>	1	Spo	-	Taphrinomycotina
<i>Neurospora crassa</i>	1	Ncr	XP-322273	Sordariomycetes
<i>Magnaporthe grisea</i>	1	Mgi	-	
<i>Gibberella zeae</i>	1	Gze	XP_391313	
<i>Chaetomium globosum</i>	2	Cgl	-	
<i>Trichoderma reesie</i>	3	Tre	33777	
<i>Aspergillus nidulans</i>	1	Ani	XP_661409 XP_664486 XP_682638	Eurotiomycetes
<i>Aspergillus fumigatus</i>	1	Afu	EAL87152 EAL87476 XP_746468 EAL87110	
<i>Aspergillus niger</i>	1	Ang	CAD99195 ABC73717 ABC73718 ABC73719 ABC73720	
<i>Aspergillus oryzae</i>	1	Aor	BAE62045 BAE56423 BAE66408 BAE62647 BAE60526	
<i>Stagonospora nodorum</i>	2	Sno	SNU10723	Dothideomycetes
<i>Botrytis cinerea</i>	2	Bci	AAS99938	Leotiomycetes
			BC1G01599	
<i>Sclerotinia sclerotiorum</i>	2	Ssc	SS1G_08218	
Basidiomycota (complete genomes with predicted proteomes)				
<i>Cryptococcus neoformans</i>	3	Cne	-	Hymenomycetes
<i>Phanerochaete chrysosporium</i>	3	Pch	7156	
			7400	
<i>Coprinus cinereus</i>	2	Cci	CC1G_06900.1	Ustilaginomycetes
<i>Ustilago maydis</i>	1	Uma	EAK82071	
Basidiomycota (other)				
<i>Gloeophyllum trabeum</i>	4	Tra	tra_1 tra_2	
<i>Ceriporiopsis subvermispora</i>	5	Cvs	Cvs_OAH	

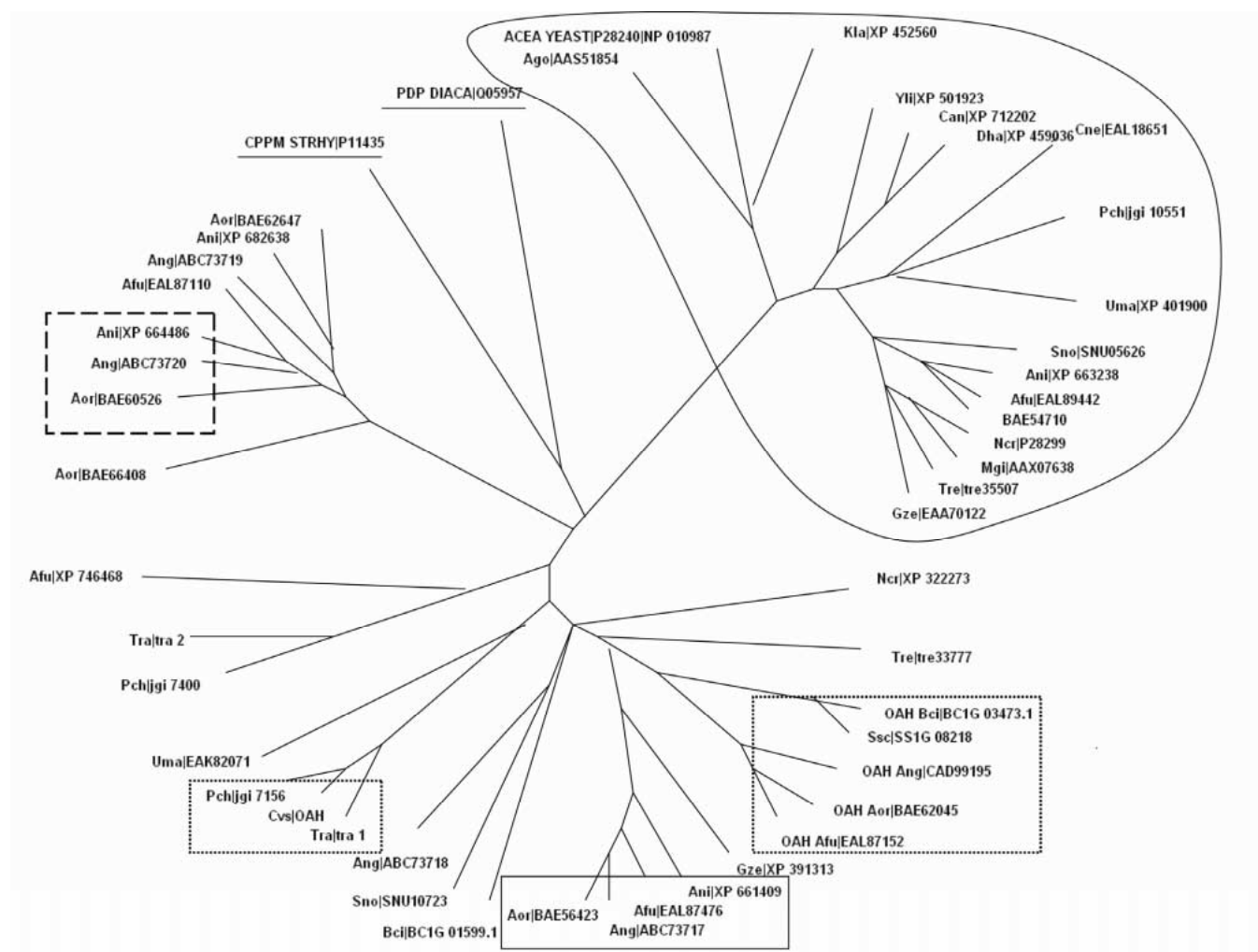
Table I OAH-like class proteins in 24 fungal proteomes. Known oxalate producers and (putative) OAH are in bold. BLAST searches were done with an expectation limit < 1e-5

\*Complete proteomes and individual sequences were retrieved from the following on line databases 1: NR-database, [www.ncbi.nlm.nih.gov/](http://www.ncbi.nlm.nih.gov/); 2: Broad institute, [www.broad.mit.edu/](http://www.broad.mit.edu/); 3: Joint Genome Institute, [www.jgi.doe.gov/](http://www.jgi.doe.gov/); 4 Concordia EST collection, <https://fungalgenomics.concordia.ca/>; 5: Sequence extracted from US Patent 6939701.

Based on BLAST searches performed using the NCBI NR database, we found that the corresponding fungal proteins belong to the isocitrate lyase/PEP mutase enzyme superfamily. The closest sequence homolog to the OAH and OAH-like proteins is the carboxyvinyl-carboxyphosphonate phosphorylmutase (CPPM, EC 2.7.8.23) of the gram-positive bacterium *Streptomyces hygroscopicus*. The closest related subclass in fungi is the methyl isocitrate lyase (MICL) subfamily present in all fungi including yeasts (Fig. 2). A third subclass of closely related proteins is represented by the *Dianthus caryophyllus* (carnation) petal death protein (PDP)<sup>21</sup>. This enzyme catalyzes the hydrolytic cleavage of oxaloacetate to oxalate and acetate as well as the C-C bond lysis of 2R-alkylmalates<sup>20</sup>.

Using the method of Folkertsma *et. al.*<sup>15</sup> a 3D-MSA of the isocitrate lyase/PEP mutase superfamily proteins was generated including all amino acid sequences from all translated orthologous and paralogous genes encoding fungal OAH-like class genes. Six most evolutionary distantly related sequences among the 18 reported X-ray structures representing the superfamily were used as starting sequences to build the alignment. These structures are isocitrate lyase (1DQU) from the fungus *Aspergillus nidulans*, two bacterial isocitrate lyases (1F61 and 1IGW), methyl isocitrate lyase (1MUM) of *Escherichia coli*, mollusk phosphoenolpyruvate phosphomutase (1PYM) and the petal death protein of *Dianthus caryophyllus* (1ZLP). The OAH-like class of fungal proteins were aligned using a profile derived from the petal death protein structure. The complete 3D-MSA is added as supplementary material or can be retrieved from our website (<http://funken.wur.nl/OAH>).

The 3D-MSA was used to construct a distance tree of OAH and OAH-like fungal proteins together with the carboxyPEP mutase of *S. hygroscopicus*, the PDP of *D. caryophyllus* and the fungal 2-methylisocitrate lyases (Fig. 2). This distance tree shows a strong clustering of the OAH proteins of *A. niger* and *B. cinerea* (both organisms are confirmed oxalate producers<sup>12,16</sup>) with putative protein EAL87152 of *A. fumigatus*, putative protein BAE62045 from *Aspergillus oryzae* and putative protein SS1G.08218 of *Sclerotinia sclerotiorum*. *A. fumigatus*<sup>17</sup>, and *S. sclerotiorum*<sup>18</sup> are also known oxalate producers. The clustering suggests that the latter three fungal genomes, as well as that of the *A. oryzae* genome, encode OAH.

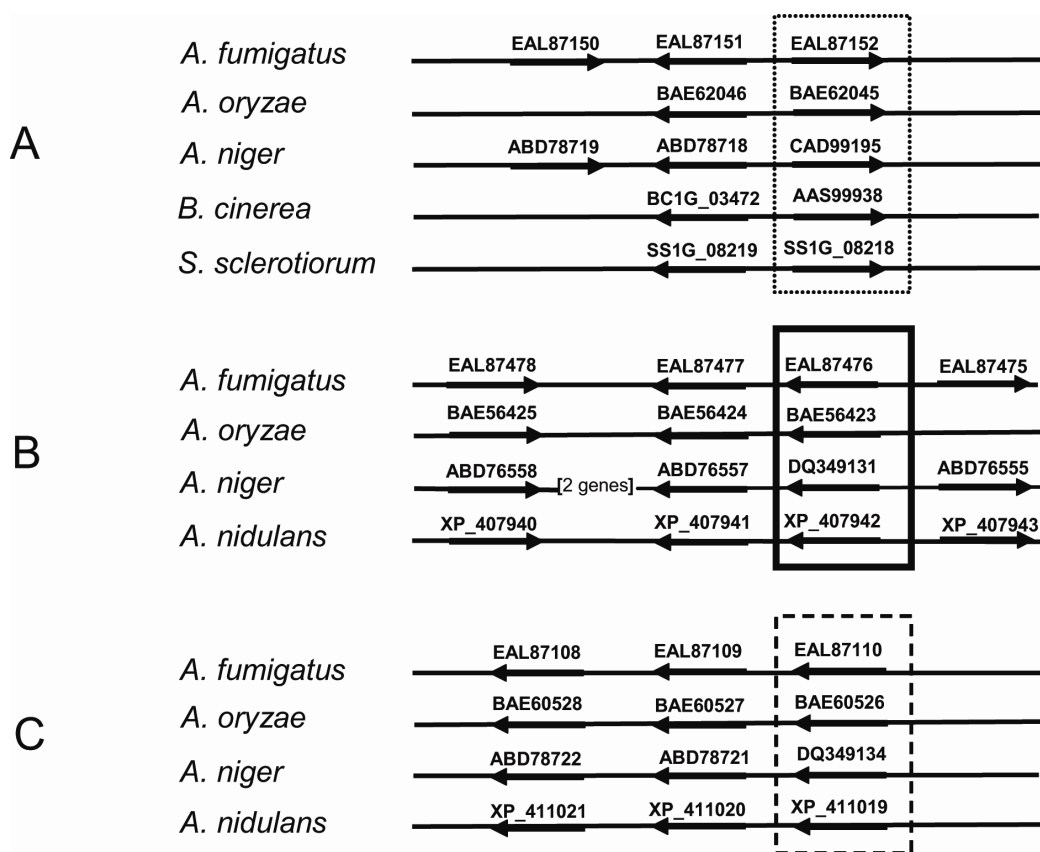


**Fig. 2** Distance tree of the OAH-like protein class, CPMP of *Streptomyces hygroscopicus*, *Dianthus caryophyllus* PDP (PDP\_IDACA) and putative fungal MICL proteins (enclosed in solid border). Rectangles enclose different orthologous groups of OAH-like class proteins. Dotted borders enclose (putative) OAH. Solid and dashed borders enclose two different orthologous groups of OAH-like class proteins

In contrast, *A. nidulans* is not an oxalate producer. Consistent with this fact is the finding that although the *A. nidulans* genome contains three OAH-like genes, the distance tree analysis suggests that none of these three genes encode OAH. This conclusion is also supported by the genome contexts in which the three OAH-like genes are found. Specifically, we screened all reported fungal genomes for possible conservation of synteny at the respective loci. Conservation of gene order was detected in the regions flanking the genes encoding the OAH of *A. niger* OAH and *B. cinerea*, and the genes encoding the putative OAHs EAL87152 of *A. fumigatus*, BAE62045 of *A. oryzae* and SS1G.08218 of *S. sclerotiorum* (Fig. 3 panel A). Conservation of gene order was also detected for the regions flanking the genes encoding the closest related paralogs of OAH, protein DQ349131 of *A. niger*, XP\_661409 of *A. nidulans* and EAL87476 of *A. fumigatus* (Fig. 3 panel B). Finally, we observed conservation of gene order in the regions flanking the genes encoding the OAH-like proteins DQ349134 of *A. niger*, EAL87110 of *A. fumigatus* and XP\_664486 of *A. nidulans* (Fig. 3 panel C). Detection of conservation of gene order at these three loci suggests that within *Aspergilli* there are at least three clusters of

orthologous groups of OAH-like proteins. No further conservation of gene order was found surrounding any other OAH class protein of any of the other fungi.

Multiple, closely related genes encoding OAH class proteins were found in the completed basidiomycete fungal genomes. The whiterot fungus *Phanerochaete chrysosporium* is an oxalate producer<sup>19</sup>. The genome encodes two OAH-like class proteins (Table I). The closest homolog of the *C. subvermispora* OAH protein is ORF 7156 suggesting that this putative *P. chrysosporium* protein is in fact OAH (Fig. 2). In addition, we detected a probable OAH encoding gene in the inferred proteome of *Coprinus cinereus* (Table I), and we could assemble from an EST sequence library from the brown-rot basidiomycete *Gloeophyllum trabeum* two unigenes encoding OAH-like class proteins. One of these is very similar (76 % sequence identity) to the *C. subvermispora* OAH.

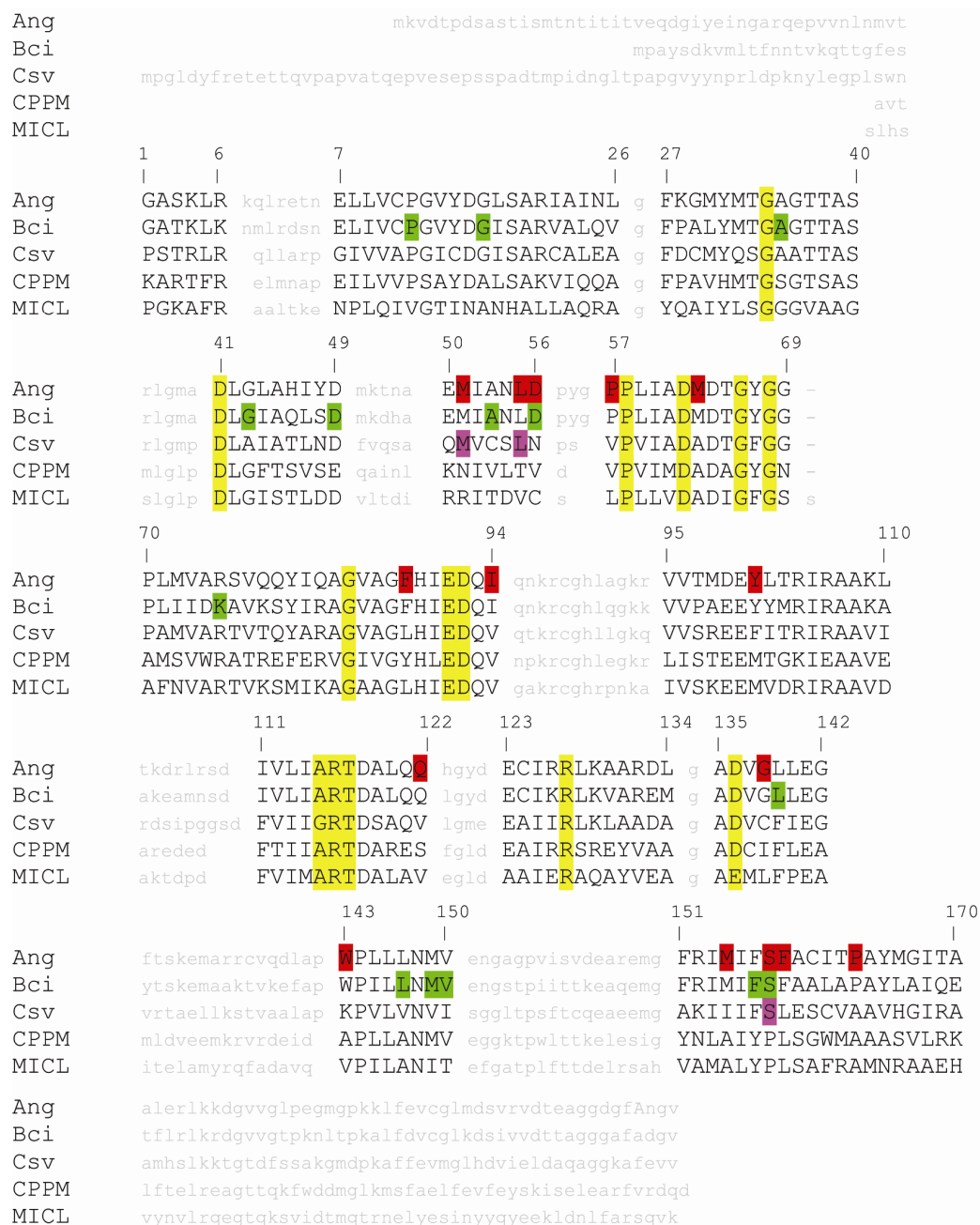


**Fig. 3 Conservation of synteny between fungal loci encoding OAH-class proteins.** Arrows represent genes. Rectangles enclose the OAH-like class protein encoding genes. Dotted borders enclose (putative) OAH. Solid and dashed borders enclose two different orthologous groups of OAH-like class proteins. Encoded protein accession numbers are indicated above the arrows. Panel A: Conservation of synteny at the OAH locus. Panel B and C: Conservation of synteny at loci encoding other OAH-like class proteins.

### 3.2 Identification of OAH Specific Residues.

The enzymes of the isocitrate lyase/PEP mutase enzyme superfamily characterized to date are known to catalyze  $Mg^{+2}$  (or  $Mn^{+2}$ )-dependent C-C bond or P-C bond forming/cleaving reactions, which proceed via  $\gamma$ -oxyanion carboxylic acid intermediates and/or transition states. The functional subfamilies of enzymes acting on C-C bonds include ketopantoate hydroxymethyl transferase (C-C bond formation), the  $\gamma$ -hydroxyacid lyases, isocitrate lyase, 2-methylisocitrate lyase, the plant petal death protein (a nonspecific lyase for C(2) substituted malates and oxaloacetate), and the fungal oxaloacetate acetylhydrolase. The subfamilies of enzymes catalyzing reactions of P-C bonds includes PEP mutase and carboxyPEP mutase, which function in phosphonate biosynthesis, and phosphonopyruvate hydrolase, an enzyme which functions in phosphonate degradation. With the exceptions of the fungal OAH and carboxyPEP mutase, each subfamily is represented by the X-ray structure of one or more liganded enzymes. Based on these structures we know that the catalytic scaffold of the isocitratelase/PEP mutase superfamily is formed at the C-terminal edge of an  $\alpha$ -barrel by nine peptide segments, one of which is derived from a swapped C-terminal  $\alpha$ -helix of an adjacent subunit. The other eight peptide segments are derived from the C-terminal regions of the  $\alpha$ -strands, the loops connecting the  $\alpha$  and  $\beta$ -elements of the barrel, and the N-terminus of a non-barrel helix. Some residues of the catalytic scaffold are conserved throughout the superfamily (these are core residues used in binding the metal ion cofactor and the substrate  $\gamma$ -oxyacid unit) while others vary to form the catalytic machinery tailored for the reaction catalyzed by a particular functional subfamily (these are diversification residues). Our objective was to use the isocitrate lyase/PEP mutase enzyme superfamily 3D-MSA that we had developed to identify an OAH specific residue(s).

Accordingly, the isocitrate lyase/PEP mutase enzyme superfamily 3D-MSA was screened for positions that are more than 90% conserved in the more than 400 unique sequences of the alignment. Matching this query are G34, D41, P58, D62, G66, G68, G84, E91, D92, A115, R116, T117, R127 and D136 (3D- numbering scheme) (Fig. 4 and Fig. 5 panel A). Except for A115, which is a glycine in the *C. subvermispora* OAH, these residues are completely conserved in all members of the fungal OAH-like protein class. Many of these positions are located around the active site near the metal ion cofactor and the substrate  $\gamma$ -oxyacid group (Fig. 5 panel B). Next, functionally related diversification residues were identified. This was accomplished by scanning the 3D-MSA for positions that show co-evolution. Residues at positions that show co-evolution tend to be conserved. However, when they do change, a group wise substitution pattern is observed reflecting the different functions of different members. Correlated mutation analysis can reveal these networks of co-evolving positions<sup>22</sup>.

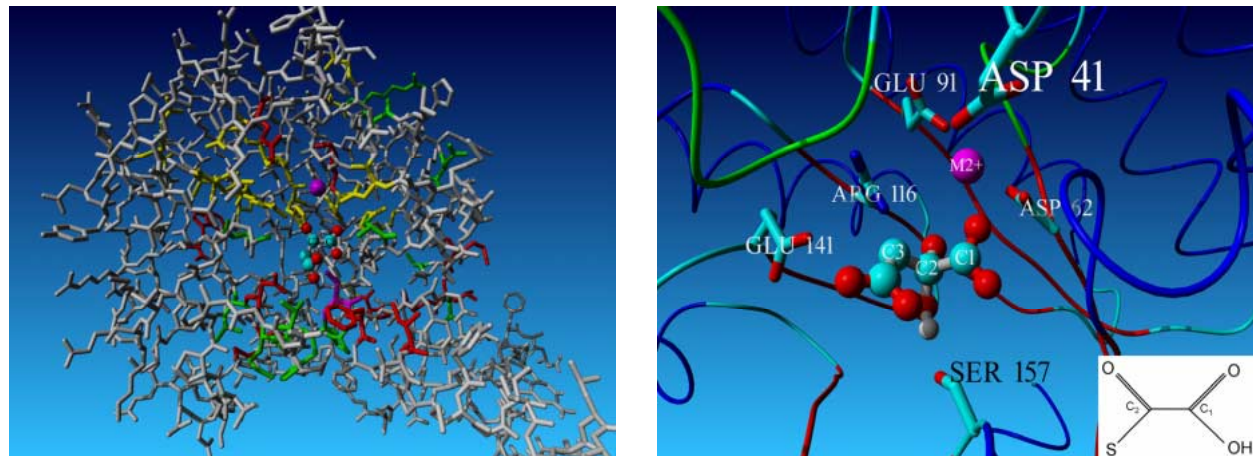


**Fig. 4: Excerpt of the structure based multiple sequence alignment of the PEP superfamily.** Included in this subset are OAH of *A. niger*, *B. cinerea* and *C. subvermispora* together with CPPM of *S. hygroscopicus* and MICL of *E. coli*. (For accession numbers see Table I) Alignment positions that are structurally conserved within the PEP superfamily are in capitals. Structurally variable parts (lower case, grey) should be considered as not aligned. Positions that are more than 90% conserved within the complete superfamily are highlighted yellow. Alignment positions that highly correlate with at least three other alignment positions are highlighted green and are indicated in the *B. cinerea* sequence. The 15 alignment positions that are 100% conserved within the four ascomycete OAH protein sequences are highlighted red and are indicated in the *A. niger* sequence. These residues are present in less than 5% of the remaining sequences. Of these 15 positions three positions are evolutionary conserved within all in (predicted) OAH sequences; these are highlighted purple in the *C. subvermispora* sequence. Note that only S157 is conserved in all (predicted) OAH proteins and also part of the network of highly correlating positions.



Fifteen alignment positions that showed high correlation with at least 3 other alignment positions were identified (Fig. 4 and Fig. 5 panel A). These residues are mainly associated with the region of the catalytic scaffold that houses the substrate moiety that is attached to the -oxyacid unit. In the isocitrate and 2-methylisocitrate lyases, this region binds the succinyl moiety that is eliminated from the substrate as a result of the C-C bond cleavage. In the petal death protein, this region binds the acetate region that is eliminated from oxaloacetate and from the 2-alkylmalate substrates, and in the organophosphonate metabolizing enzymes this region interacts with the substrate phosphonate substituent.

We then screened the 3D-MSA for positions that are 100% conserved within the known OAHs from ascomycete fungi and at such a position the corresponding amino acid was found in less than 5% of the remaining sequences. This screen resulted in a set of 15 OAH specific residues (Fig. 4). Of these 15 alignment positions only alignment position 157 was also shown to be part of the network of highly correlated positions that separate the different functions present in the superfamily. Alignment position 157 is a serine in all proteins with OAH activity including the petal death protein (S257). Importantly, this serine residue is not found in any of the other OAH-like class proteins. Inspection of the published active site of the petal death protein modeled with (2R, 3S)-2-ethyl-3-methylmalate<sup>21</sup> reveals that the petal death protein S257 (3D-number S157) is directed at the 2-ethyl group.



**Fig. 5: Conserved and functionally related diversification residues in the 3D model structure of oxaloacetate hydrolase from *Aspergillus niger*.**

*Panel A (left): Positions that show more than 90% conservation within the PEP superfamily alignment are in yellow. Positions that show a high CMA score are in green. OAH specific residues are in red. Amino acid S157 (in purple) is specific for OAH and shows a high CMA score.*

*Panel B (right): Gem-diol of oxaloacetate modeled in the active site of petal death protein (1ZLP). Side-chains of important active site residues conserved in the superfamily are shown (white label). Ser157 important for OAH activity (black label) is optimally positioned for hydrogen bonding the gem-diol. M2+ indicates the metal ion involved in the catalysis. Insert: The  $\alpha$ -keto-acid group, a common backbone of the substrates of the PEP superfamily enzymes.*

By modeling the gem-diol of oxaloacetate in the active site the petal death protein we found that the pro-R C(2)OH is positioned for deprotonation (which sets off the C-C bond cleavage) and the pro-S C(2)OH (corresponding to the 2-ethyl substituent of the (2R, 3S)-2-ethyl-3-methylmalate ligand), is positioned for hydrogen bond interaction with the S257 side chain. Based on the knowledge that oxaloacetate equilibrates with the gem-diol in aqueous solution, and that the petal death protein catalysis C(2)-C(3) bond cleavage in 2R-alkylmalates, it has been hypothesized that the petal death protein converts oxaloacetate to oxalate and acetate by first binding the gem-diol form of oxaloacetate (a minor component of the equilibrium with the ketone) and then subjecting it to the same catalytic cycle used in lysis of the 2R-alkylmalate substrates<sup>20</sup>. An alternative mechanism involves addition of an active site water molecule at the C(2)=O of the bound oxaloacetate to form the C(2)gem-diol intermediate, followed by C(2)-C(3) bond cleavage. The same reaction mechanism appears to be operative in OAH, evidenced by the ability of OAH to catalyze the cleavage (albeit slowly) (3S, 2R)-2,3-dimethylmalate, in addition the hydrolytic cleavage of oxaloacetate. We have recently found that the C(2) gemdiol of 3,3-difluorooxaloacetate binds to OAH with high affinity, which is consistent with strong binding interactions between the enzyme active site and the oxaloacetate C(2) gem-diol in the enzyme-substrate complex or in an enzyme-intermediate complex.<sup>23</sup> The conservation of S157 (3D-numbering scheme) among OAH active proteins thus appears to have a functional rather than a structural basis. Namely, S157 might bind the pro-S C(2)OH of the oxaloacetate gem-diol substrate through hydrogen bond formation. This interaction would contribute favorable binding energy for substrate binding and/or catalysis. To test this hypothesis we carried out the mutagenesis study described in the following section.

### 3.3 Evaluation of the S157 OAH Sequence Marker.

The contribution that the active site serine makes in catalysis of the cleavage of oxaloacetate to oxalate and acetate was evaluated in the *B. cinerea* OAH and in the carnation petal death protein. Residues with 3D-number 157 of *B. cinerea* OAH (S260) and of petal death protein (S257) were mutated to Ala, Pro and Thr so to replicate the residue usage in the other known isocitrate lyase/PEP mutase family members. The four OAH-like proteins of *A. niger* contain Pro at this position. The effect of the Ser replacement on the kinetic constants for catalysis of oxaloacetate conversion to oxalate and acetate can be seen from inspection of Table II.

Enzyme	$k_{\text{cat}}$ ( $\text{s}^{-1}$ )	$K_{\text{m}}$ ( $\mu\text{M}$ )	$k_{\text{cat}}/K_{\text{m}}$ ( $\text{M}^{-1}\text{s}^{-1}$ )
OAH-WT	$17.4 \pm 0.2$	$65 \pm 3$	$2 \times 10^5$
OAH-S260T <sup>a</sup>	$2.59 \pm 0.06$	$7000 \pm 400$	$4 \times 10^2$
OAH-S260T <sup>b</sup>	$3.6 \pm 0.1$	$7700 \pm 500$	$5 \times 10^2$
OAH-S260A <sup>a</sup>	$0.83 \pm 0.01$	$700 \pm 30$	$1 \times 10^3$
OAH-S260A <sup>b</sup>	$0.71 \pm 0.01$	$1307 \pm 60$	$5 \times 10^2$
OAH-S260P <sup>a</sup>	$0.96 \pm 0.03$	$150 \pm 20$	$6 \times 10^3$
OAH-S260P <sup>b</sup>	$0.98 \pm 0.03$	$220 \pm 20$	$5 \times 10^3$
PDP-WT	$2.72 \pm 0.06$	$130 \pm 10$	$2 \times 10^4$
PDP-S257T <sup>a</sup>	$1.14 \pm 0.02$	$3600 \pm 200$	$3 \times 10^2$
PDP-S257T <sup>b</sup>	$1.16 \pm 0.03$	$2600 \pm 200$	$5 \times 10^2$
PDP-S257A	$2.91 \pm 0.07$	$1530 \pm 80$	$2 \times 10^3$
PDP-S257P	$4.1 \pm 0.1$	$840 \pm 50$	$5 \times 10^3$

**Table II Steady-state kinetic constants determined for OAH and PDP.** Catalyzed conversion of oxaloacetate to oxalate and acetate in 5 mM MgCl<sub>2</sub> and 0.1 M imidazole buffer (pH=7.5, 25 °C). <sup>a</sup> Trial 1, <sup>b</sup> Trial 2.

The  $k_{\text{cat}}$  decreases by a factor of only 5 to 20. In contrast the  $K_{\text{m}}$  values change more than 100 fold for the S157T mutation. When taking the  $k_{\text{cat}}$  values into consideration, it is clear that the Ser to Thr replacement is best tolerated. This might be due to the ability of the Thr hydroxyl group to engage in hydrogen bond formation with the C(2)OH of the oxaloacetate gem-diol. On the other hand, the  $k_{\text{cat}}/K_{\text{m}}$  determined for the Thr mutant is considerably lower than that of the wild-type OAH (~1000-fold) and the S157A or S157P OAH mutants. The apparent drop in substrate binding affinity is probably attributed to an unfavorable steric interaction introduced by the methyl group of the Thr secondary alcohol. The S157A and S157P mutants display  $k_{\text{cat}}$  values that are 3-fold less than that of the S157T mutant and 20-fold less than that of the wild-type OAH. This decrease in  $k_{\text{cat}}$  values relative to the S157T mutant may be due to the absence of hydrogen bond formation to the Pro-S C(2)OH of the oxaloacetate C(2) gem-diol. However, in contrast to the S157T mutant the substrate binding is less influenced in the S157A and S157P mutant, probably due to a diminished steric interaction in the latter two relative to the S157T mutant. The  $k_{\text{cat}}/K_{\text{m}}$  value of the S157A mutant is 300-fold less than that of wild-type OAH, whereas that of the S157P mutant is 50-fold less. Based on these results, it is clear that the OAH S157 makes a significant contribution to OAH catalysis.

To determine if S257 (3D-number S157) of the petal death protein also makes a significant contribution to catalysis of oxaloacetate cleavage to oxalate and acetate, it too was replaced with Thr, Ala and Pro. The steady-state kinetic constants measured for oxaloacetate cleavage are listed in Table II; whereas, the  $k_{\text{cat}}$  value was not significantly altered by the S257 replacements, the  $k_{\text{cat}}/K_{\text{m}}$  was noticeably decreased: ~100 fold for the S257T substitution and ~10-fold for the S257A or S257P substitution. This effect is not as substantial as that observed with OAH, however this is to be

expected in view of the fact that the petal death protein active site has evolved to function as a scavenger for C(2) substituted malates. The OAH active site by comparison is specialized in oxaloacetate cleavage. Indeed, not only is the  $k_{\text{cat}}/K_m$  for oxaloacetate cleavage 10-fold greater than that of the petal death protein, but its substrate specificity is much greater (Table II); whereas the petal death protein can catalyze the cleavage of the 2R-ethyl-3S-methylmalate with the same efficiency as the oxaloacetate, the OAH catalyzed lysis reaction of (2R, 3S) 2,3-dimethylmalate is 100-fold slower than the hydrolytic cleavage of the oxaloacetate (Tables II, III, IV).

Enzyme	$k_{\text{cat}}$ ( $\text{s}^{-1}$ )	$K_m$ ( $\mu\text{M}$ )	$k_{\text{cat}}/K_m$ ( $\text{M}^{-1}\text{s}^{-1}$ )
OAH-WT	$0.0293 \pm 0.0007$	$140 \pm 10$	$2 \times 10^2$
OAH-S260A	$0.0249 \pm 0.0002$	$54 \pm 2$	$5 \times 10^2$
OAH-S260P	$0.074 \pm 0.002$	$22 \pm 2$	$3 \times 10^3$

**Table III. Steady-state kinetic constants determined for the OAH.** Catalyzed conversion of 2R, 3S-2,3-dimethylmalate to pyruvate and propionic acid in 5 mM  $\text{MgCl}_2$  and 50 mM HEPES buffer (pH=7.5, 25 °C).

WT or Mutants	$k_{\text{cat}}$ ( $\text{s}^{-1}$ )	$K_m$ ( $\mu\text{M}$ )	$k_{\text{cat}}/K_m$ ( $\text{M}^{-1}\text{s}^{-1}$ )
PDP-WT	$8.4 \pm 0.1$	$530 \pm 30$	$2 \times 10^4$
PDP-S257A	$5.4 \pm 0.2$	$560 \pm 50$	$1 \times 10^4$
PDP-S257P	$8.3 \pm 0.2$	$660 \pm 40$	$1 \times 10^4$

**Table IV. Steady-State Kinetic Constants Determined for the PDP.** Catalyzed conversion of 2R-ethyl-3S-methylmalate to  $\alpha$ -ketobutyrate and propionic acid in 5 mM  $\text{MgCl}_2$  and 50 mM HEPES buffer (pH=7.5, 25 °C).

It is well a well-known fact among mechanistic enzymologists that the replacement of an enzyme active residue is likely to reduce catalytic efficiency ( $k_{\text{cat}}/K_m$ ), independent of whether the residue directly interacts with the reactant. For this reason, we could not unequivocally conclude that the OAH and petal death active site Ser residues assume the role that we have proposed simply based on reduced activity in the Ser mutants. To arrive at such a conclusion, we needed an internal control to determine whether the Ser substitution had altered to configuration of active site residues. The control experiment was carried out to determine the impact of the Ser mutation on catalysis of C-C bond cleavage in an analogous substrate that lacked the pro S C(2)OH of the oxaloacetate gem-diol. Accordingly, we measured the kinetic constants for wild-type OAH and OAH Ser mutants catalysis of C-C bond cleavage in 3S,2R-dimethylmalate (the most active substrate following oxaloacetate<sup>23</sup>), and the kinetic constants for wild-type petal death protein and petal death protein Ser mutants catalyzed C-C bond cleavage in 3S-methyl, 2R-ethyl-malate (the most active substrate). The results are shown in Table III and IV, respectively. In contrast to the reduction observed in efficiency of OAH catalysis of oxaloacetate cleavage upon Ser260 (3D-number S157) replacement, the efficiency of catalysis of C-C bond cleavage in 3S,2R-dimethylmalate actually improved slightly. In the case of the petal death protein the catalytic efficiency was essentially unchanged, consistent with the comparatively lax substrate specificity exhibited by this particular lyase.

### 4. Methods

#### 4.1 Enzymes.

Recombinant *Botrytis cinerea* OAH and *Dianthus caryophyllus* petal death protein were prepared according to published procedure<sup>20,23</sup>. Site-directed mutants were prepared using a PCR based strategy with *pfu* polymerase. The plasmids OAH-pET-3c and petal death protein-Pet-3c served as template in conjunction with commercial custom primers. Gene sequences were verified by nucleotide sequencing. The mutant proteins were purified to homogeneity (as judged by SDS-PAGE analysis) using the same protocol reported for the purification of the wild-type enzymes<sup>20,23</sup>. The mutant proteins behaved chromatographically the same as the wild-type enzymes and were stable upon storage.

#### 4.2 Activity Assays.

The conversion of oxaloacetate to oxalate and acetate was monitored using the assay as reported<sup>20</sup>. The lysis reactions of (3*S*, 2*R*) 2-ethyl-3-methylmalate and (3*S*, 2*R*) 2,3-dimethylmalate were monitored as described previously<sup>20</sup>.

#### 4.3 Steady-State Kinetic Constant Determination.

The steady-state kinetic parameters ( $K_m$  and  $k_{cat}$ ) for reactions catalyzed by OAH and the petal death protein were determined from the initial velocity data measured as a function of substrate concentrations. The initial velocity data were fitted to Equation 1 with KinetAsystI

$$V_0 = V_{max}[S]/(K_m + [S]) \quad (1)$$

where [S] is the substrate concentration,  $V_0$  is the initial velocity,  $V_{max}$  is the maximum velocity, and  $K_m$  is the Michaelis-Menten constant for the substrate. The  $k_{cat}$  value was calculated from  $V_{max}$  and the enzyme concentration using the equation  $k_{cat} = V_{max}/[E]$ , where [E] is the protein subunit molar concentration.

#### 4.4 Data collection, creation and screening of the 3D-MSA.

The Isocitrate lyase/PEP mutase enzyme superfamily protein structures were collected from the Structural Classification of Proteins (SCOP)<sup>24</sup> structure comparison resource. The superfamily

## Chapter 3

sequences were collected by BLAST searches using the sequences of the structures as queries. To collect all closely related fungal proteins from the 18 annotated publicly available fungal genomes BLAST searches were performed in the corresponding protein sets. To find additional OAH like proteins BLAST searches were performed in the Concordia EST collection database (<https://fungalgenomics.concordia.ca/>). The 17 structure files of the superfamily were divided into subfamilies according to SCOP. Structure superposition and sequence alignment were performed using WHATIF<sup>25</sup>. The complete superfamily alignment was built as described<sup>15</sup>. The alignment was used for correlated mutation analyses (CMA) using a CMA-score algorithm based on the method described by Shulman *et. al.*<sup>26</sup> Details of the algorithm can be found at our website (<http://3dmiscis.systemsbiology.nl/corMutIdx.php>). OAH specific amino acids were identified by scanning the alignment for positions that are 100% conserved in all OAHs and a different residue in the rest of the alignment. Different is here defined as present in less than 5% of the remaining sequences.

### 4.5 Fungal phylogenetic tree.

Twelve fungal panorthologous genes were selected and the protein sequences were aligned using clustalW<sup>27</sup>. Alignments were manually curated for ambiguous aligned sequences and concatenated. Maximum likelihood phylogenetic analysis was carried out with Tree Puzzle<sup>28</sup> using the VT model<sup>29</sup> and a gamma model of rate heterogeneity with alpha= 0.64

### 4.6 Distance tree.

A neighbor joining tree was built of all collected proteins belonging to the OAH-like protein class, together with CPPM from *S. hygroscopicus* and the putative methyl isocitrate lyase (MICL) proteins from all 17 completely sequenced fungi functionally related diversification residues used in this study. Only the structurally conserved regions were used as aligned input for QuickTree<sup>30</sup> to build the tree. The tree was visualized with Treeview<sup>31</sup>.

### 4.7 Analysis of synteny.

Two ORFs directly adjacent to ORFs encoding OAH-like class proteins were used to query fungal genomes encoding one or more OAH-like class proteins by blastX and if a single very highly significant hit resulted (E-value < 1E-10), this was considered to be the likely ortholog of the query. When two genes were found to be adjacent in two different species, they were determined to be located in the same synteny segment in those species

### *4.8 Modeling of gem-diol bound in the active site of PDP.*

1ZLP was used as template to create a structure of PDP with bound gem-diol of oxaloacetate. The gem-diol was created using the build option of YASARA followed by energy minimization (EM) using YASARA's YAMBER2 force field as described<sup>32</sup>. The resulting gem-diol was modeled in the active site of PDP by superimposing it on 5-hydroxypentanal: a compound covalently bound in the active site of PDP in structure 1ZLP. A hydroxyl group attached to C2 (Fig. 5B) of the gem-diol was superimposed on the hydroxyl group of 5-hydroxypentanal. C2 and the next two carbons (R-group, Fig. 5B) of the gem-diol were superimposed on the first three carbon atoms attached to the hydroxyl group of 5-hydroxypentanal. After superpositioning, the remaining hydroxyl group attached to C2 in the gem-diol was optimally positioned for hydrogen bond formation with S157 (3D-numbering scheme) (Fig. 5B).

## 5. Conclusions

The isocitrate lyase/PEP mutase enzyme superfamily is represented in fungal genomes by isocitrate lyase, 2-methylisocitrate lyase and one or more members of the OAH-like class of proteins. In *A. niger* for example the OAH class contains OAH encoded by *oahA* and two genes encoding enzymes that share 52 % sequence identity with the OAH and two genes encoding enzymes that share 34% sequence identity with the OAH. Each of these five enzymes contain catalytic residues that distinguish the C-C bond lyase branch of the family, namely the Cys general base and Glu general acid. Our working hypothesis is that all are in fact C-C bond lyases, however only one is specialized for oxaloacetate cleavage. The specialization derives from a strategically placed serine residue S281 (3D number S157) which functions to bind and orient the minor, hydrated form (gem-diol) of oxaloacetate. Presently, we do not know the exact nature of the catalytic functions performed by the other four enzymes. In each of these sequences a proline occupies alignment position 157, which implies that they are in fact not OAH active enzymes. This is consistent with genetic studies that have demonstrated that oxalate production in *A. niger* is totally dependent on the OAH encoding gene<sup>12,14,23</sup>. In on going work we aim to identify the functions of these 4 OAH homologs.

The results from our studies are evidence that the following uncharacterized proteins are OAHs: EAL87152 of *A. fumigatus*, BAE62045 of *A. oryzae*, SS1G\_08218 of *S. sclerotiorum*, the protein encoded by ORF 7156 of *Phanerochaete chrysosporium*, CC1G\_06900.1 of *Coprinus cinereus*, and OAH of *G. trabeum*. It is noteworthy that although basidiomycete fungi are evolutionary very distant from the ascomycete fungi (Fig. 1) the S157 is still conserved (Fig. 4) and that all known oxalic acid producers possess a predicted OAH encoding gene (Table I) (Fig. 2), consistent with the view that the conversion of oxaloacetate into oxalate and acetate is the main route for oxalic acid biosynthesis in fungi.



## 6. References

1. Dutton, Evans CS. Oxalate production by fungi: its role in pathogenicity and ecology in the soil environment. *Can J Microbiol* 1996;42:881-895.
2. Kirkland BH, Eisa A, Keyhani NO. Oxalic acid as a fungal acaricidal virulence factor. *J Med Entomol* 2005;42:346-351.
3. Maxwell DP, Bateman DF. Oxalic acid biosynthesis by *Sclerotium rolfsii*. *Phytopathology* 1968;58:1635-1642.
4. Godoy G, Steadman JR, Dickman B, Dam R. Use of mutants to demonstrate the role of oxalic acid in pathogenicity of *Sclerotinia sclerotiorum* on *Phaseolus vulgaris*. *Physiol Mol Plant Pathol* 1990;37:179-191.
5. Nakagawa Y, Shimazu K, Ebihara M, Nakagawa K. *Aspergillus niger* pneumonia with fatal pulmonary oxalosis. *J Infect Chemother* 1999;5:97-100.
6. Kubicek CP, Schreierl-Kunar G, Wohrer W, Rohr M. Evidence for a cytoplasmic pathway of oxalate biosynthesis in *Aspergillus niger*. *Appl Environ Microbiol* 1988;54:633-637.
7. Balmforth AJ, Thomson A. Isolation and characterization of glyoxylate dehydrogenase from the fungus *Sclerotium rolfsii*. *Biochem J* 1984;218:113-118.
8. Hammel KE, Mozuch MD, Jensen KA Jr, Kersten PJ. H<sub>2</sub>O<sub>2</sub> recycling during oxidation of the arylglycerol beta-aryl ether lignin structure by lignin peroxidase and glyoxal oxidase. *Biochemistry* 1994;33:13349-13354.
9. Yaver D, Cherry B, Murrell J. Polypeptides having oxaloacetate hydrolase activity and nucleic acids encoding same. 2004 Patent 6939701.
10. Munir E, Yoon gem JJ, Tokimatsu T, Hattori T, Shimada M. A physiological role for oxalic acid biosynthesis in the wood-rotting basidiomycete *Fomitopsis palustris*. *Proc Natl Acad Sci USA* 2001;98:11126-11130.
11. Akamatsu Y, Takahashi M, Shimada M. Influences of various factors on oxaloactase activity of the brown-rot fungus *Tyromyces palustris*. *Mokuzia Gakkaishi* 1993;39:352-356.
12. Ruijter GJ, van de Vondervoort PJ, Visser J. Oxalic acid production by *Aspergillus niger*: an oxalate-non-producing mutant produces citric acid at pH 5 and in the presence of manganese. *J Microbiology* 1999;145:2569-2576.
13. Pedersen H, gem C, Nielsen J. Cloning and characterization of oah, the gene encoding oxaloacetate hydrolase in *Aspergillus niger*. *J Mol Gen Genet* 2000;263:281-286.
14. Pedersen H, Christensen B, Hjort C, Nielsen J. Construction and characterization of an oxalic acid nonproducing strain of *Aspergillus niger*. *Metab Eng* 2000;2:34-41.
15. Folkertsma S, van Noort P, Van Durme J, Joosten HJ, Bettler E, Fleuren W, Oliveira L, Horn F, de Vlieg J, Vriend G. A family-based approach reveals the function of residues in the nuclear receptor ligand-binding domain. *J Mol Biol*. 2004;341:321-335.
16. Verhoeff K, Leeman M, Peer R, Posthuma L, Schot N, van Eijk GW. Changes in pH and the production of organic acids during colonisation of tomato petioles by *Botrytis cinerea*. *J Phytopathol* 1988;122:327-336.
17. Khabir A, Makni S, Ayadi L, Boudawara T, Frikha I, Sahnoun Y, Jlidi R. Pulmonary oxalosis with necrotizing pulmonary aspergillosis. *Ann Pathol* 2002;22:121.
18. Guimaraes RL, Stotz HU. Oxalate production by *Sclerotinia sclerotiorum* deregulates guard cells during infection. *Plant Physiol* 2004;136:3703-3711.
19. Kenealy WR, Dietrich DM. Growth and fermentation responses of *Phanerochaete chrysosporium* to O<sub>2</sub> limitation. *Enzyme and microbial technology* 2004;34:490-498.
20. Lu Z, Feng X, Song L, Han Y, Kim A, Herzberg O, Woodson WR, Martin BM, Mariano PS, Dunaway-Mariano D. Diversity of function in the isocitrate lyase enzyme superfamily: the *Dianthus caryophyllus* petal death protein cleaves alpha-keto and alpha-hydroxycarboxylic acids. *Biochemistry* 2005;44:16365-16376.
21. Teplyakov A, Liu S, Lu Z, Howard A, Dunaway-Mariano D, Herzberg O. Crystal structure of the petal death protein from carnation flower. *Biochemistry* 2005;44:16377-16384.
22. Oliveira L, Paiva CM, Vriend G. Correlated mutation analyses on very large sequence families. *Chembiochem* 2002;3:1010-1017.
23. Han Y, Joosten HJ, Niu W, Zhao Z, Mariano PS, McCalman MT, van Kan JAL, Schaap PJ, Dunaway-Mariano D. Oxaloacetate hydrolase: the c-c bond lyase of oxalate secreting fungi. *J Biol Chem* 2007;282:9581-9590.
24. Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res* 2004;32:226-229.
25. Vriend G. WHAT IF: a molecular modeling and drug design program. *J Mol Graph* 1990;8:52-56

## Chapter 3

26. Shulman AI, Larson C, Mangelsdorf DJ, Ranganathan R. Structural determinants of allosteric ligand activation in RXR heterodimers. *Cell* 2004;116:417–429.
27. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994;22:4673-4680.
28. Schmidt HA, Strimmer K, Vingron M, von Haeseler A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 2002;18:502-504.
29. Muller T, Vingron M. Modeling amino acid replacement. *J Comput Biol* 2000;7:761-776.
30. Howe K, Bateman A, Durbin R. QuickTree: building huge Neighbour-Joining trees of protein sequences. *Bioinformatics* 2002;18:1546-1547.
31. Page RD. TreeView: an application to display phylogenetic trees on personal computers. *Comput Appl Biosci* 1996;12:357-358.
32. Van Durme J, Horn F, Costagliola S, Vriend G, Vassart G. GRIS: Glycoprotein-hormone Receptor Information System. *Mol Endocrinol*. 2006; 20:2247-2255.

## Chapter 4

Oxaloacetate hydrolase: The C-C bond lyase of oxalate  
secreting fungi

## Oxaloacetate hydrolase: The C-C bond lyase of oxalate secreting fungi.

### 1 Abstract

Oxalate secretion by fungi is known to be associated with fungal pathogenesis. In addition, oxalate toxicity is a concern for the commercial application of fungi in the food and drug industries. Although oxalate is generated through several different biochemical pathways, oxaloacetate acetylhydrolase (OAH) catalyzed hydrolytic cleavage of oxaloacetate appears to be an especially important route. Below, we report the cloning of the *Botrytis cinerea oahA* gene and the demonstration that the disruption of this gene results in the loss of oxalate formation. In addition, through complementation we have shown that the intact *B. cinerea oahA* gene restores oxalate production in an *Aspergillus niger* mutant strain, lacking a functional *oahA* gene. These observations clearly indicate that oxalate production in *A. niger* and *B. cinerea* is solely dependent on the hydrolytic cleavage of oxaloacetate catalyzed by OAH. In addition, the *B. cinerea oahA* gene was overexpressed in *E. coli* and the purified OAH was used to define catalytic efficiency, substrate specificity and metal ion activation. These results are reported along with the discovery of the mechanism-based, tight-binding OAH inhibitor 3,3-difluorooxaloacetate ( $K_i = 68$  nM). Finally, we propose that cellular uptake of this inhibitor could reduce oxalate production.

### 2 Introduction

Numerous filamentous fungi, including the food biotechnology fungus *Aspergillus niger*, the opportunistic human pathogen *Aspergillus fumigatus*, the phytopathogenic fungi *Botrytis cinerea* and *Sclerotinia sclerotiorum*, as well as many brown-rot and white-rot basidiomycetes, are able to efficiently produce large quantities of oxalate<sup>1,2</sup>. It is known that oxalate secretion is associated with fungal pathogenesis<sup>1,3-6</sup>. In the wood-rotting fungus *Fomitopsis palustris* oxalate is formed as the product of glucose metabolism<sup>7</sup>. We recently initiated investigations of the oxalate biosynthetic pathway in order to develop a genomic-based method for distinguishing between oxalate producing and non-producing fungi. An additional goal of this effort was to identify enzyme inhibitors that could be used to arrest oxalate formation in targeted fungi.

In order to attenuate oxalate production in fungi, it is necessary to first identify the major pathway responsible for oxalate formation. There are three potential routes for production of oxalate in fungi: oxidation of glyoxylate<sup>8,9</sup>, oxidation of glycolaldehyde<sup>10</sup>, and hydrolysis of oxaloacetate<sup>11</sup>. The results of studies of [14C]CO<sub>2</sub> incorporation into the metabolite pools of *A. niger* indicate that oxalate is derived from oxaloacetate<sup>12</sup>. This finding parallels the results of earlier work on the purification of an

enzyme "oxalacetalase" (now known as oxaloacetate acetylhydrolase or OAH) that catalyzes the hydrolytic cleavage of oxaloacetate to form acetate and oxalate<sup>11</sup>. In a subsequent study, a mutant *A. niger* strain, NW228<sup>13</sup>, was found to be deficient in both oxalate production and in the synthesis of active OAH<sup>14</sup>. These observations suggest that oxalate is produced only by the OAH catalyzed process. In the present investigation, we verified the connection between the absence of oxalate formation and the absence of OAH activity in the NW228 mutant by demonstrating that the OAH-encoding *oahA* gene is interrupted by a stop codon. In an independent effort, Pedersen *et al.*<sup>15</sup> mutated the *A. niger oahA* gene by recombination with an *oahA* sequence-based plasmid, to create a metabolically robust *A. niger* strain deficient in oxalate production.

OAH isolated from *A. niger* is reported to be a mixture of N-terminal truncates<sup>16</sup>. To obtain homogeneous OAH, we first cloned the *A. niger oahA* gene for overexpression in *E. coli*. However, we failed to isolate active transformants. Subsequent efforts focused on cloning the *oahA* gene from the alternate fungal source *Botrytis cinerea* (previously known as *Botryotinia fuckeliana*). The cloned *B. cinerea oahA* gene was shown to both restore oxalate production in the *A. niger* NW228 mutant strain (13, 14) that lacks a functional *oahA* gene and produce OAH in transformed *E. coli* cells. In order to provide insight into the mechanism of catalysis and to set the stage for the design of mechanism-based inhibitors, the OAH substrate and metal cofactor specificities were determined. Below we report the results from these studies, which have culminated in the discovery that 3,3-difluorooxaloacetate is a novel tight-binding OAH inhibitor.

### 3. Materials and methods

#### 3.1 Materials

Oxaloacetic acid, (*R*)-malic acid, (*S*)-malic acid, (2*R*)-methylmalic acid, (2*S*)-methylmalic acid, (2*R*, 3*S*)-isocitric acid were purchased from Sigma-Aldrich. 2*R*, 3*S*-2, 3-dimethylmalic acid, 2*R*-ethyl-3*S*-methylmalic acid, 2*R*-propyl-3*S*-methylmalic acid, *threo*-(2*R*,3*S* & 2*S*,3*R*)-2-methylisocitrate were prepared according to reference 17. (2*R*, 3*S*)-isopropylmalate (synthesized by M. Jung) and 3-butylmalate (Aldrich Rare Chemicals, catalog no. S789046) were provided by S. Clarke (University of California, Los Angeles, CA).

3,3-Difluorooxaloacetate was prepared by modifying the protocol described by Saxty *et al.*<sup>18</sup>. A stirred suspension of zinc powder (6.5 g, 0.1 mol) in anhydrous tetrahydrofuran (50 mL) containing Me<sub>3</sub>SiCl (4 mL, 0.04 mol) was stirred at room temperature for 15 min and then at reflux while ethyl bromodifluoroacetate (24.24g, 0.12mol) was added. To this solution was added ethyl formylformate (5.2 g, 0.05 mol) at a rate sufficient to maintain gentle reflux. After stirring at reflux for 1 h, the solution was cooled to room temperature and concentrated *in vacuo*. The resulting residue was dissolved in water and then extracted with ethyl acetate. The extracts were washed with water, dried

and concentrated *in vacuo*. The resulting residue was subjected to silica gel column chromatography (methylene chloride and 1:5 ethyl acetate/hexane) to give diethyl 2,2-difluoro-3-hydroxysuccinate (7.9 g, 69 %).

To a solution of sulfuric acid (10 mL, 0.18 mol) and pyridinium chlorochromate (0.093 mol) in water (40 mL) was added drop-wise diethyl 2,2-difluoro-3-hydroxysuccinate (7.0 g, 0.031 mol). After stirring at room temperature for 7 h, the mixture was filtered, The filtrate was concentrated *in vacuo* to give a residue, which was subjected to silica gel column chromatography (10:1 to 3:1 ethyl acetate/hexane) to give diethyl 3,3-difluorooxaloacetate (3.75g, 50%). The spectroscopic data of this substance matched those reported earlier<sup>19</sup>. <sup>1</sup>H NMR (CDCl<sub>3</sub>) 1.34 (m, 6H), 4.39 (m, 4H); <sup>13</sup>C NMR 167.1, 162.3 (t, J = 31.0 Hz), 110.8 (t, J = 263.5 Hz), 91.9 (t, J = 28.1 Hz), 64.1, 63.7, 13.7.

A mixture of diethyl 3,3-difluorooxaloacetate (1.70 g, 7.0 mmol) and 13 mL of 3N HCl was stirred at reflux for 1.3 h, cooled to room temperature, and concentrated *in vacuo*. Addition of 2 mL TFA to the residue resulted in the formation of a solid, which was filtered and dried *in vacuo* giving 3,3-difluorooxaloacetate (1.18 g, 90 %). In agreement with a previous report<sup>19</sup>, the <sup>13</sup>C NMR (D<sub>2</sub>O) spectrum is as follows: 169.9, 165.4 (t, J = 29.4 Hz), 112.0 (t, J = 261.1 Hz), 92.0 (t, J = 27.2 Hz).

### 3.2 *A. niger* OAH Gene Cloning and Sequencing

The *oahA* genes from the *A. niger* wild type strain N400 (CBS120.49), and derived mutant strains NW228 (*prtF28*) and NW229 (*prtF29*)<sup>13</sup>, were amplified by using standard PCR protocol in conjunction with the forward primer CTGGCCCTTCCTTTCTATC and the reverse primer CCATCCAATGCAGTTCAAC. The PCR products were cloned using the vector pGEM®-T Easy (Promega) and sequenced. The nucleotide sequence of the N400 strain has been deposited in the public databases under accession no AJ567910.

### 3.3 Cloning of the *B. cinerea* *oahA* Gene

The GenomeWalker kit (clontech laboratories Inc.) was used to obtain the PCR product of the genomic region containing the OAH encoding gene from *B. cinerea* strain B05.10. The sequences of gene specific primer 1 (ATCAACACAATATCGGAGTTCATGG) and primer 2 (GCACGAATTCTCATGTAG-TACTCCTCT) were derived from *B. cinerea* EST AL117176. The complete *oah* gene and 308 bp of the upstream region was amplified and cloned in the vector pGEM®-T Easy (Promega). and sequenced. Standard RT-PCR techniques were used to verify the two introns. The nucleotide sequence has been deposited in the public databases under accession no AY590264.

### 3.4 Complementation of the *A. niger prtF* Mutation

Three different constructs were made for complementation of the *prtF* mutation with the *B. cinerea oah* gene. These are the full length *B. cinerea oah* gene including the 308 bp upstream region and two fusion constructs in which the *A. niger pkiA*<sup>20</sup> promoter was fused to the two possible start codons. All three constructs were co-transformed into strain NW188 (*pyrA6, prtF28*) as previously described<sup>21</sup>. Transformants were screened for oxalate production by growing single colonies on complete media (CM)<sup>22</sup> plates with 10 ng/ml methyl orange as pH indicator. Only the transformants in which the *pkiA* promoter was fused to the most upstream start codon showed oxalate production as verified by HPLC analysis. The copy number of OAH *B. cinerea* of the transformants were determined by Southern analysis using the *pkiA* promoter as a radiolabeled probe. The copy number was estimated by comparing the intensity of the native *A. niger pki* promoter band (1 copy) and the intensity of the band of the introduced *pkiA-BcoahA* construct.

### 3.5 Targeted Mutagenesis of *Botrytis cinerea*

A gene replacement construct containing fragments originating from either end of the *BcoahA* gene flanking a hygromycin resistance cassette (Genbank accession AJ439603) was constructed. The 5'-flanking fragment (466 bp) was amplified using primers CCCAATCCTCCAAGAGAAGTC and GATTACTAACAGATATCAAGGCTTCAAGCGGGAAGCAGTGGTAC. The 3'-flanking fragment (611 bp) was amplified using primers GGGTACCGAGCTGCAATT-CGTTGTGGACATCTCCAAGGC and CCAACCAGGTACTGAGATCAG. The flanking fragments were joined to the hygromycin cassette by overlap extension PCR in a single reaction with three template fragments and primers GACTGCTACTGAGTATTC-GGT and CTACTCAAACACCATCCGCGA. The resulting PCR fragment was excised from the gel and directly transformed to *B. cinerea* protoplasts as described in reference<sup>23</sup>.

### 3.6 Purification of Recombinant *Botrytis cinerea* OAH from the OAH-pET3c *E. coli* Clone

The OAH gene was amplified by using a PCR-based strategy<sup>24</sup> with the OAH-pGem-T easy vector clone serving as the template. *Pfu* DNA polymerase (Stratagene) and oligonucleotide primers containing *NdeI* and *BamHI* restriction sites were used for the subcloning the OAH gene into the pET-3c vector (Novagen). The recombinant plasmid, OAH-pET3c, was used to transform competent *E. coli* BL21(DE3) cells (Novagen). The transformed cells were grown at 20 °C with mild agitation (180 rpm) in Luria broth (LB) containing 50 µg/ml carbenicillin. After 19 h of cell growth (OD<sub>600nm</sub> ~ 0.7), induction was initiated with 0.4 mM IPTG (RPI Corp.). The culture was incubated for 12.5 h at 20 °C under conditions of vigorous mixing (200 rpm). The cells were harvested by centrifugation (6500

RPM (7,808g)) for 15 min at 4°C in a yield of 4 g/L of culture media. The cell pellet (23 g) was suspended in 230 ml of ice-cold lysis buffer (50 mM K<sup>+</sup> Hepes (pH 7.5), 1 mM EDTA, 1 mM benzamide hydrochloride, 0.05 mg/ml trypsin inhibitor, 1 mM 1, 10-phenanthroline, 0.1 mM PMSF and 5 mM DTT). The suspension was passed through a French Press at 1200 PSI, and then centrifuged at 4 °C for 60 min at 20000 RPM (48,384g). The supernatant was fractionated by ammonium sulfate induced protein precipitation. The 30-40% ammonium sulfate protein precipitate was dialyzed at 4 °C against 4 x 2 L Buffer A (50 mM triethanolamine (pH 7.5), 5 mM MgCl<sub>2</sub> and 5 mM DTT) before loading onto a 4.5 x 45 cm DEAE-cellulose column equilibrated with 2 L Buffer A. The column was washed with 1 L of Buffer A, and then eluted with a 2 L linear gradient of 0 to 0.3 M KCl in Buffer A. The column fractions were analyzed by measuring the absorbance at 280 nm and by carrying out SDS-PAGE analysis. Ammonium sulfate was added to the combined fractions to generate 20 % saturation. The resulting solution was loaded onto a 3 x 30 cm Butyl Sepharose column equilibrated at 4 °C with 20 % ammonium sulfate in 500 mL of Buffer A. The column was washed with 450 mL of 20 % ammonium sulfate in Buffer A and then eluted with a 1 L linear gradient of 20 % to 0 % ammonium sulfate in Buffer A. The column fractions were analyzed by measuring the absorbance at 280 nm and by carrying out SDS-PAGE analysis. The OAH eluted at 4 % ammonium sulfate. The OAH-containing fractions were combined, concentrated at 4 °C with an Amicon concentrator (Amicon) and then dialyzed against Buffer A. The resulting sample was concentrated using a MACROSEP 10K OMEGA for storage at -80 °C. The protein sample was shown to be homogeneous by SDS-PAGE analysis. Yield: 4 mg protein/g wet cell.

### *3.7 Recombinant Botrytis cinerea OAH N-Terminal Sequence Determination*

OAH was chromatographed on a SDS-PAGE gel, transferred to a PVDF membrane (Novex Co.) and subjected to automated protein N-terminal amino acid sequencing by Dr. Brian M. Martin of the National Institutes of Mental Health (Molecular Structure Unit, Department of Neurotoxicology, Bethesda Maryland, USA) to obtain the sequence PAYSDKVMLT.

### *3.8 Recombinant Botrytis cinerea Molecular Mass Determination*

The theoretical subunit molecular mass of recombinant OAH was calculated from the amino acid composition, derived from the gene sequence, by using the EXPASY Molecular Biology Server program Compute pI/MW<sup>25</sup>. The subunit size of recombinant OAH was determined by SDS-PAGE analysis with molecular weight standards from Invitrogen. The subunit mass was determined by MS-ES mass spectrometry (University of New Mexico Mass Spectrometry Lab). The molecular weight of native recombinant OAH was determined using gravity flow gel filtration techniques. The chromatography of OAH was carried out on a 1.5 x 180 cm Sephacryl S-200 column (Pharmacia)



## Chapter 4

equilibrated with Buffer B (25 mM K<sup>+</sup>Hepes, 0.15 M KCl, 0.5 mM DTT, pH 7.5) at 4 °C that had been calibrated using Pharmacia Gel Filtration Calibration Kit (catalase 232,000; aldolase 158,000; albumin 67,000; Ovalbumin 43,000; chymotrypsinogen A 25,000; Ribonuclease A 13,000) according to the manufacture's instructions. The chromatography was carried out at 4 °C using Buffer B as eluant and a peristaltic pump to maintain a constant flow rate of 1 mL/min. The plots of the elution volume of the molecular weight standards vs log molecular weight were found to be linear. The molecular weight was thus derived from the measured elution volume by extrapolation.

### 3.9 Oxaloacetate Hydrolase Assay

OAH activity was assayed according to a published procedure<sup>11</sup>. Reaction solutions (1 mL) contained 0.06-1.0 mM oxaloacetate, 0.032 μM OAH and 5 mM MgCl<sub>2</sub> or 0.18-1.2 mM oxaloacetate, 0.011 μM OAH and 0.3 mM MnCl<sub>2</sub> in 0.1 M imidazole (pH 7.6 and 25 °C). The reaction was monitored at 255 nm for the disappearance of the enol tautomer of oxaloacetate ( $\Delta\epsilon = 1.1 \text{ mM}^{-1} \text{ cm}^{-1}$ ). The rate of oxaloacetate consumption via spontaneous decarboxylation was measured prior to initiating the enzymatic reaction in order to determine the “background rate”. The background rate was subtracted from the reaction rate measured in the presence of OAH. The influence of the buffer properties on the OAH kinetic behavior was tested by replacing the 0.1 M imidazole of the reaction solutions with 50 mM K+HEPES (pH 7.5) or 50 mM Tris-HCl (pH 7.5).

### 3.10 Assay for Malate Substrates

**Continuous Assay:** The OAH lyase activity towards 1mM (*R*)-malate, (*S*)-malate, 2*S*-methylmalate, (2*R*, 3*S*)-isocitrate, *threo*-(2*R*,3*S* & 2*S*,3*R*)-2-methylisocitrate, (2*R*,3*S*)-isopropylmalate and 3-butylmalate was measured using 1 mL reaction solutions containing 5 mM MgCl<sub>2</sub>, 20 units/mL LDH, and 0.2 mM NADH in 50 mM K+Hepes (pH 7.5 and 25 °C). The absorbance of the solution was monitored at 340 nm ( $\Delta\epsilon = 6.2 \text{ mM}^{-1} \text{ cm}^{-1}$ ). The kinetic constants for OAH lysis of (2*R*)-methylmalate, (2*R*, 3*S*)-dimethylmalate, and (2*R*)-ethyl-(3*S*)-methylmalate were determined with assay solutions in which the reactant concentration is varied from 0.5Km to 10Km. In the case of 2*R*-ethyl-3*S*-methylmalate, reaction solutions contained 600 units of LDH.

**Fixed-Time Assay:** Reaction solutions initially containing 6 μM OAH, 0.50-3.75 mM (2*R*)-propyl-(3*S*)-methylmalate, 5 mM MgCl<sub>2</sub>, and 50 mM K+Hepes (pH 7.5, 25 °C) were analyzed at ~20 % conversion. A 200 μL aliquot was mixed with 200 μL of 1 N HCl and 100 μL of 0.4 M phenylhydrazine hydrochloride. The solution was stirred for 12 min before the absorbance was measured at 326 nm (2-oxovaleric acid hydrazone molar extinction constant is  $6.8 \text{ mM}^{-1} \text{ cm}^{-1}$ ). The control reaction lacked OAH.

### 3.11 Steady-State Kinetic Constant Determination for Recombinant *Botrytis cinerea* OAH Substrates

The steady-state kinetic parameters ( $K_m$  and  $k_{cat}$ ) for reactions catalyzed by OAH were determined from the initial velocity data measured as a function of substrate concentrations. The initial velocity data were fitted to Equation 1 with KinetAsystI

$$V_0 = V_{max}[S]/(K_m + [S]) \quad (1)$$

where  $[S]$  is the substrate concentration,  $V_0$  is the initial velocity,  $V_{max}$  is the maximum velocity, and  $K_m$  is the Michaelis-Menten constant for the substrate. The  $k_{cat}$  value was calculated from  $V_{max}$  and the enzyme concentration using the equation  $k_{cat} = V_{max}/[E]$ , where  $[E]$  is the protein subunit molar concentration in the reaction calculated from the ratio of measured protein concentration and the protein molecular mass (34486 Da).

### 3.12 Determination of Inhibition Constants for Recombinant *Botrytis cinerea* OAH Inhibitors

The competitive inhibition constant  $K_i$  was determined by measuring the initial velocity of product formation at 25 °C as a function of substrate and inhibitor (I) concentrations and fitting the initial velocity data to equation 2 with KinetAsystI. For determination of the difluorooxaloacetate  $K_i$  the 1 mL reaction solutions initially contained 0.1 M imidazole (pH 7.6), 5 mM  $MgCl_2$ , 9 nM oxaloacetate hydrolase, varying concentrations of oxaloacetate (0.08-1.0 mM) and changing, fixed concentrations of difluorooxaloacetate (0, 0.2, 0.4  $\mu$ M). Reactions were monitored at 255 nm as described<sup>11</sup>. This same protocol was used in the determination of the inhibition constants for oxalate (0, 25, 50 and 100  $\mu$ M), *R*-malate (0, 10 and 20 mM) and *S*-malate (0, 20 and 40 mM).

$$V_0 = V_{max} [S]/(K_m(1+(I/K_i)) + [S]) \quad (2)$$

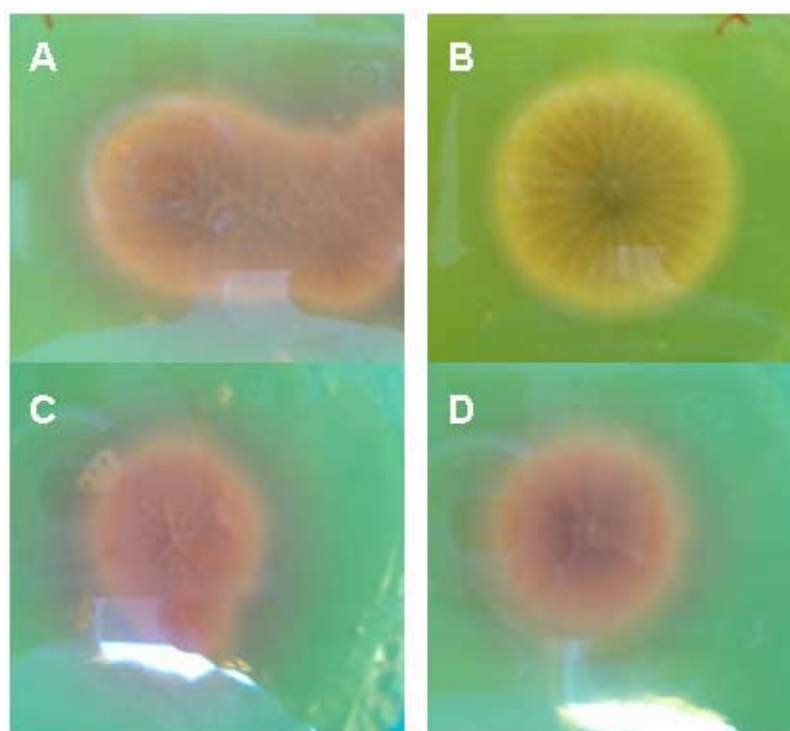
### 3.13 Metal Ion Activation of Recombinant *Botrytis cinerea* OAH

The metal ion-free protein was prepared by exhaustive dialysis at 4 °C against 30 mM EDTA/50 mM triethanolamine/5mM DTT (pH 7.5) in 7 × 500 mL (2 h each time), and then against 50 mM triethanolamine /5 mM DTT (pH 7.5) in 6 × 500 ml (2 h each time). The initial velocity of oxaloacetate consumption in reaction solutions containing OAH, 1 mM oxaloacetate, and varying concentrations of  $MgCl_2$ ,  $MnCl_2$ ,  $CoCl_2$ ,  $CaCl_2$ ,  $ZnCl_2$ ,  $FeSO_4$ ,  $CuBr_2$ ,  $BaCl_2$ ,  $NiCl_2$  in 0.1 M imidazole buffer (pH 7.6, 25°C) was measured by monitoring the absorbance change at 255 nm<sup>11</sup>. The initial velocity data were analyzed using Equation 1 and the computer program KinetAsystI.

## 4. Results

### 4.1 Cloning and sequencing of the *oahA* gene from oxalate non-producing mutant strains

The *oahA* genes from the *A. niger* mutant strains, *prtF28* (NW228) and *prtF29* (NW229) were sequenced and their sequences were compared to that of the *oahA* gene contained in the wild-type parental strain. Previous studies had shown the mutant strains lack both oxalate production and OAH activity<sup>14,16</sup>. However it remained to be demonstrated that the *oahA* gene is disrupted. The sequence analysis verified that the *oahA* genes from the mutant strains had acquired a stop codon in the open reading frame (ORF) (at amino acid position R94 (NW228) or Q159 (NW229)). Furthermore, by using a pH indicator plate assay (Figure 1) and HPLC analyses (data not shown) we demonstrated that the mutant strains, transformed with the wild type *oahA* gene, regain their oxalate production capability.



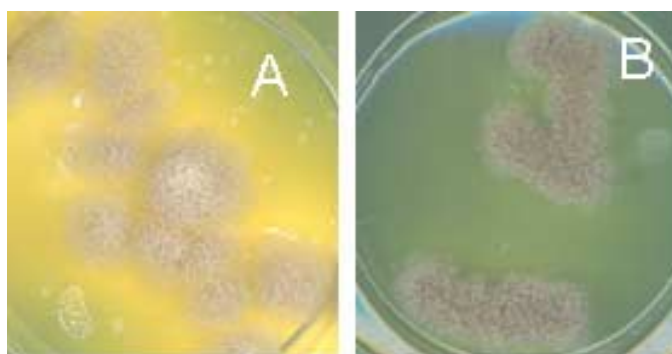
**Fig. 1 Screen for oxalate production *A. niger* using a methyl orange pH-indicator plate assay.** The pink color indicates acidification caused by oxalate production. Panel A: wild type *A. niger*. Panel B: oxalate non-producing mutant NW228. Panel C: NW228 transformed with the wild type *A. niger oahA* gene. Panel D: NW228 transformed with the *B. cinerea oahA* gene.

### 4.2 Cloning and disruption of the *Botrytis cinerea oahA* gene

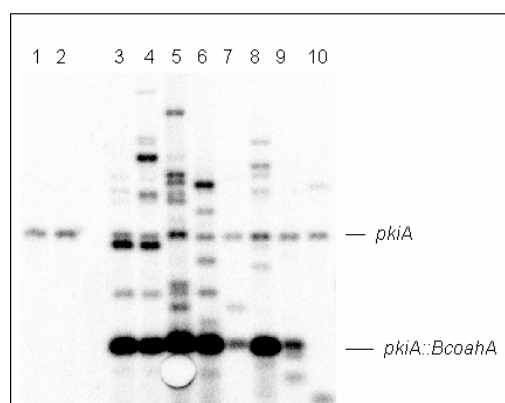
The results from the complementation experiment described above strongly suggest that oxalate production in *A. niger* is caused solely by OAH catalyzed cleavage of oxaloacetate. We scrutinized the genomes of other fungal oxalate producers to see if they too encoded OAH. BLAST searches using the

*A. niger oahA* gene (*AnoahA*) as the query suggested that EST AL117176 from the oxalate-producer *Botrytis cinerea* represents a partial copy of the *B. cinerea* OAH encoding gene. The complete *BcoahA* gene, isolated from the *B. cinerea* genome by using PCR and genome walking techniques, was then used for complementation of the *prtF28* mutation. The sequence of the putative BcOAH was shown to be 70% identical to that of the *A. niger* OAH. Because there are two possible start codons separated by 7 amino acids, three different constructs were made for complementation purposes. These are the full length *B. cinerea oahA* gene (including the 308 bp upstream region calculated from the most upstream candidate start codon) and two fusion constructs, in which the constitutive *A. niger pkiA* promoter<sup>20</sup> is fused to one of the two possible start codons. Complementation of the *A. niger* NW228 non-oxalate producing strain was achieved with the *BcoahA* coding region that included the most upstream start codon fused to the strong *A. niger pkiA* promoter<sup>20</sup>. Seven independent *pkiA::BcoahA* transformants were identified by using an indicator plate assay (Figure 2). The integration of the construct within the transformants was verified by using Southern analysis (Figure 3). HPLC analysis was used to demonstrate that each transformant had regained the ability to produce oxalate (data not shown).

To confirm the role played by the *BcoahA* gene in oxalate biosynthesis in *B. cinerea*, the gene was deleted from the *B. cinerea* genome by employing targeted gene replacement. Transformants, in which homologous recombination had occurred, were identified by PCR screening and by Southern blot analysis. Five independent transformants were identified as having perfect gene replacement in the absence of additional ectopic integration (data not shown). The pH indicator plate assay (Figure 2) was used to demonstrate that the *BcoahA*-deficient mutants do not produce a detectable level of oxalate.



**Fig. 2 Screen for oxalate production *B. cinerea* using a bromothymol blue pH-indicator plate assay.** The yellow color indicates acidification caused by oxalate. Panel A: wild type *B. cinerea*. Panel B: *oahA* deletion strain.



**Fig. 3 Molecular characterization of *A. niger* NW228 transformants complemented with *pkiA::BcoahA* construct.** Lane 1 and 2 NW228 control. Lane 3–9: eight independent NW228 *pkiA::BcoahA* transformants. Lane 10: Wild type control. The bands corresponding with the native *pkiA* gene and the construct are indicated.

#### 4.3 Purification of Recombinant *B. cinerea* OAH and Size Determination

The DNA sequence of the subcloned gene was found to agree with the published sequence (GenBank™ accession number AY590264) except that the nucleotide at position 1637 is G not A. Thus, the encoded amino acid is Ala not Thr. The recombinant OAH was purified to homogeneity in an overall yield of 4 mg/g wet cells by using the 4-step protocol summarized in Table 1.

Purification step	Total protein (g)	Total activity <sup>a</sup> (units)	Specific activity (units/mg)	Yield (%)	Purification (-fold)
Cell extract	20	1400	0.070	100	
Ammonium Sulfate	4	1200	0.28	85	4
DEAE-Cellulose	0.34	940	2.7	68	39
Butyl-Sepharose	0.17	520	3.1	38	44

**Table 1 Experimental protocol for purification of *Botrytis cinerea* OAH from *E. coli* BL21(DE3) cells transformed with the OAH-pET3c clone**

<sup>a</sup> One unit of enzyme activity is defined as the amount of enzyme required for the consumption of 1  $\mu$ mol oxaloacetate/min in 0.1M imidazole, 5 mM  $MgCl_2$ , and 1 mM oxaloacetate at 25 °C and pH 7.6.

The N-terminal sequence of OAH revealed that the N-terminal methionine is lost during post-translational modification. This was confirmed by a molecular mass determination by using mass spectroscopy. The theoretical mass of OAH-Met is 34355 Da compared to the experimental value of  $34355 \pm 1$  Da. The SDS-PAGE analysis gave an estimated subunit mass of 35 kDa, whereas the native mass measured by using molecular size gel filtration chromatography is ~100 kDa. This result is indicative of a homotrimeric quaternary structure. The  $\alpha,\beta$ -barrels of the enzymes of the PEP mutase/isocitrate lyase superfamily possess the C-terminal  $\alpha$ -helix from an adjacent subunit, a family

structural trait known as ‘helix swapping’. The functional unit is therefore a dimer, and the quaternary structure that has been most frequently encountered is the homotetramer<sup>17</sup>. Because a trimeric structure is not consistent with this pattern, we examined the possibility that the association state of OAH is unstable. Native molecular mass determination, by using size exclusion chromatography, coupled with on-line laser scattering, refractive index, and ultraviolet detection provided additional information regarding the OAH quaternary structure. A sample of OAH eluted in a single peak from size exclusion fractionation was polydispersed with a MW range of 50 kDa at 0.05 mg/ml to 114 kDa at 0.3 mg/mL and averaged MW of 97 kDa. At 1.1 mg/ml the sample reached an averaged MW of 118 Da and a hydrodynamic radius of 4.1 nm. The hydrodynamic radius remained at 4.1 nm at 3 mg/ml indicating that the protein does not exist in oligomeric forms higher than a tetramer. The results are consistent with the existence of a monomer-dimer-tetramer equilibrium.

#### 4.4 Metal Ion Specificity of *B. cinerea* OAH

Enzymes of the isocitrate lyase/PEP mutase superfamily require a divalent metal ion for catalysis. The steady-state kinetic constants for metal ion activation of OAH were determined at a saturating concentration of oxaloacetate (1 mM) and varying metal ion concentration using 0.1 M imidazole (pH 7.6), 50 mM K<sup>+</sup>Hepes (pH 7.5) or 50 mM Tris-HCl (pH 7.5) as buffer. The  $k_{cat}$  and  $K_a$  values, measured for the Mg<sup>2+</sup>, Mn<sup>2+</sup>, and Ca<sup>2+</sup> activation are listed in Table 2. Mn<sup>2+</sup> and Mg<sup>2+</sup> are significantly better activators for OAH than is Ca<sup>2+</sup>. The  $k_{cat}$  values determined using the 0.1 M imidazole buffer do not differ significantly from those measured using the K<sup>+</sup>Hepes buffer. The  $K_a$  values measured using the K<sup>+</sup>Hepes buffer (which does not bind divalent metal ions) are slightly smaller than those measured using the imidazole buffer (which does bind metal ions). OAH is not subject to metal ion inhibition by Mg<sup>2+</sup> (no inhibition observed at 5 mM MgCl<sub>2</sub>). Likewise, inhibition was not observed at 0.3 mM MnCl<sub>2</sub>. The metal ions Zn<sup>2+</sup> (100 μM), Co<sup>2+</sup> (300 μM), Fe<sup>2+</sup> (100 μM), Cu<sup>2+</sup> (500 μM) and Ni<sup>2+</sup> (1 mM) failed to activate OAH.

Metal activator	$K_m$ (μM)	$k_{cat}$ (s <sup>-1</sup> )
Mg <sup>2+</sup>	94 ± 9	10.1 ± 0.3
Mn <sup>2+</sup>	40 ± 3	40 ± 1
Ca <sup>2+</sup>	730 ± 90	0.70 ± 0.03

**Table 2** Steady-state kinetic constants for metal ion activators of *Botrytis cinerea* OAH catalyzed hydrolytic cleavage of 1 mM oxaloacetate in 0.1 M imidazole (pH 7.6, 25°C). The metal ions Zn<sup>2+</sup> (100 μM), Co<sup>2+</sup> (300 μM), Fe<sup>2+</sup> (100 μM), Cu<sup>2+</sup> (500 μM) and Ni<sup>2+</sup> (1 mM) failed to activate OAH

4.5 Substrate Specificity and Kinetic Properties of *B. cinerea* OAH

The steady-state kinetic constants for OAH catalyzed hydrolysis of oxaloacetate were measured as a function of oxaloacetate concentration at fixed metal ion cofactor concentration (5 mM MgCl<sub>2</sub> or 0.3 mM MnCl<sub>2</sub>). Reactions were carried out at 25 °C using 0.1 M imidazole (pH 7.6), 50 mM K<sup>+</sup>Hepes (pH 7.5) or 50 mM Tris-HCl (pH 7.5) as buffer. The results are summarized in Table 2. Mn<sup>+2</sup> activated OAH displays a higher turnover rate than does the Mg<sup>+2</sup> activated OAH. On the other hand, the K<sub>m</sub> of oxaloacetate measured by using Mn<sup>+2</sup> as activator is larger than that measured with Mg<sup>+2</sup> serving as the activator. Consequently, there is no significant difference in the specificity constant  $k_{\text{cat}}/K_{\text{m}}$  between Mg<sup>+2</sup> and the Mn<sup>+2</sup> activated OAH.

The ability of OAH to catalyze C-C bond cleavage in  $\alpha$ -hydroxycarboxylate metabolites was probed in order to determine if OAH retains the C-C lyase activity characteristics of the C-C bond lyase branch of the isocitrate lyase/PEP mutase family (*viz.* petal death protein, isocitrate lyase and 2-methylisocitrate lyase)<sup>17</sup>. The results are listed in Table 3. The first set of substrates tested are the *R*- and *S*-enantiomers of malate. To enhance the sensitivity for detection of product formation, reaction mixtures containing high concentrations of OAH (13  $\mu$ M) were used. In addition, to ensure saturation of catalytic sites high concentrations (1 mM) of the reactants were employed. Under these conditions, no C-C lyase activity is observed, suggesting that the  $k_{\text{cat}}$  for cleavage of malate is less than  $1 \times 10^{-5} \text{ s}^{-1}$ .

Next, substrate activities of C(2) alkyl malates were probed. Whereas the (2*S*)-methylmalate is not a substrate for OAH, the (2*R*)-methylmalate is converted to pyruvate and acetate with a  $k_{\text{cat}} = 0.01 \text{ s}^{-1}$  and a  $K_{\text{m}} = 1.45 \text{ mM}$  by this enzyme. The addition of a methyl group at the C(3) position of the (2*R*)-methylmalate leads to a further improvement in substrate activity. The  $K_{\text{m}}$  value of (2*R*, 3*S*)-dimethylmalate is 10-fold smaller than that for (2*R*)-methylmalate. However, the  $k_{\text{cat}}$  values measured for these two substrates are equivalent. The substrate (2*R*)-ethyl-(3*S*)-methylmalate displayed a 10-fold larger  $k_{\text{cat}}$  and a comparably larger  $K_{\text{m}}$ . The  $k_{\text{cat}}/K_{\text{m}}$  value measured for (2*R*)-propyl-(3*S*)-methylmalate is significantly smaller ( $49 \text{ M}^{-1}\text{s}^{-1}$ ), a finding that suggests that the active site has limited space for the C(2) alkyl group.

The native substrates of isocitrate lyase ((2*R*, 3*S*)-isocitrate) and 2-methylisocitrate lyase ((2*R*, 3*S* & 2*S*,3*R*)-2-methylisocitrate) are not substrates for OAH. This observation indicates that the OAH active site cannot accommodate a C(3) CH<sub>2</sub>COO- substituent.

In summary, (2*R*, 3*S*)-2,3dimethylmalate and (2*R*)-ethyl-(3*S*)-methylmalate are the most active of the malate substrates tested, each having a  $k_{\text{cat}}/K_{\text{m}}$  of *ca.*  $2 \times 10^2 \text{ M}^{-1}\text{s}^{-1}$ . The petal death protein is known

(17) to possess the same substrate specificity: oxaloacetate ( $k_{\text{cat}}/K_m = 2 \times 10^4 \text{ M}^{-1}\text{s}^{-1}$ ) and (2*R*)-ethyl-(3*S*)-methylmalate ( $k_{\text{cat}}/K_m = 2 \times 10^4 \text{ M}^{-1}\text{s}^{-1}$ ) > (2*R*)-propyl-(3*S*)-methylmalate ( $k_{\text{cat}}/K_m = 3 \times 10^3 \text{ M}^{-1}\text{s}^{-1}$ ) > (2*R*)-methylmalate ( $k_{\text{cat}}/K_m = 7 \times 10^2 \text{ M}^{-1}\text{s}^{-1}$ ) >>> (*R*)-malate and isocitrate. However, the catalytic efficiencies for cleavage of these substrates by the petal death protein are considerably higher than that observed for OAH.

Reactant	$k_{\text{cat}}(\text{s}^{-1})$	$K_m(\mu\text{M})$	$k_{\text{cat}}/K_m(\text{M}^{-1}\text{s}^{-1})$
Oxaloacetate	$17.4 \pm 0.2$	$65 \pm 3$	$2.7 \times 10^5$ (a)
2 <i>R</i> -methylmalate	$(1.13 \pm 0.02) \times 10^{-2}$	$1450 \pm 60$	7.8 (b)
2 <i>R</i> , 3 <i>S</i> -dimethylmalate	$(2.93 \pm 0.07) \times 10^{-2}$	$140 \pm 10$	$2.1 \times 10^2$ (b)
2 <i>R</i> -ethyl-3 <i>S</i> -methylmalate	$(3.46 \pm 0.09) \times 10^{-1}$	$2000 \pm 100$	$1.7 \times 10^2$ (c)
2 <i>R</i> -propyl-3 <i>S</i> -methylmalate	$(3.77 \pm 0.07) \times 10^{-2}$	$770 \pm 40$	49 (d)
(2 <i>R</i> , 3 <i>S</i> )-isocitrate	$<10^{-5}$ (b)		
(2 <i>R</i> , 3 <i>S</i> & 2 <i>S</i> ,3 <i>R</i> )-2-methyisocitrate	$<10^{-5}$ (b)		
<i>R</i> -malate or <i>S</i> -malate	$<10^{-5}$ (b)		
2 <i>S</i> -methylmalate	$<10^{-5}$ (b)		
3-isopropylmalate	$<10^{-5}$ (b)		
3-butylmalate	$<10^{-5}$ (b)		

**Table 3 Steady-State Kinetic Constants Determined for the C-C Bond Cleavage Reactions Catalyzed by *Botrytis cinerea* OAH in 5 mM MgCl<sub>2</sub> and 50 mM K<sup>+</sup>Hepes buffer (pH 7.5 and 25°C). Chemical structures are shown in Chart 1.**

<sup>a</sup> The kinetic constants were determined using direct optical method.

<sup>b</sup> The kinetic constants were determined using LDH/NADH coupling assay (20 units/mL LDH).

<sup>c</sup> The kinetic constants were determined by LDH/NADH coupling assay (600 units/mL LDH).

<sup>d</sup> The kinetic constants were determined using the fixed-time phenylhydrazine-based assay.



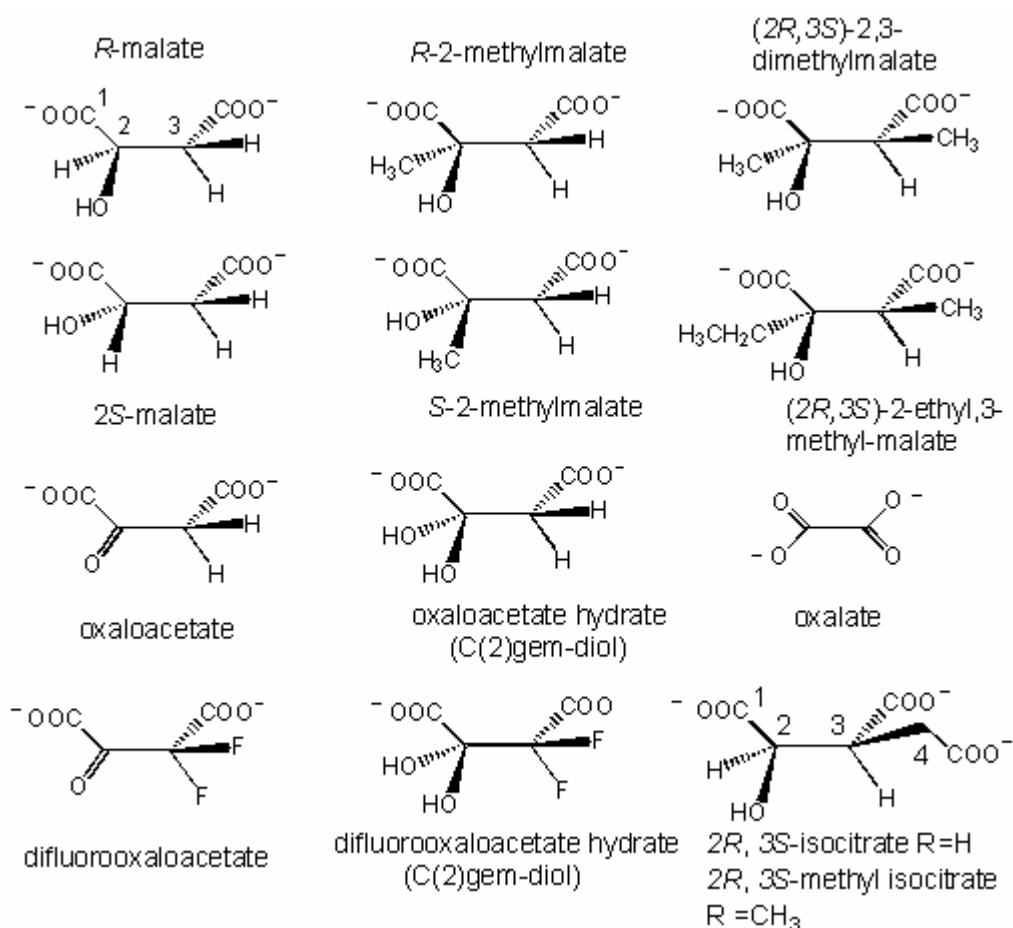


Chart 1

#### 4.6 Inhibition of *B. cinerea* OAH

The competitive inhibition constants of oxalate, (*R*)-malate, (*S*)-malate and 3,3-difluorooxaloacetate were evaluated in order to gain information about the structural determinants for ligand binding. Oxalate, a product of oxaloacetate cleavage, binds to OAH with high affinity ( $K_i = 19 \pm 1 \mu\text{M}$ ). Oxalate is also an analog of the pyruvate enolate anion intermediate formed in the cleavage of (*2R*, *3S*)-2,3-dimethylmalate. The petal death protein binds oxalate tightly ( $K_i = 4.3 \pm 0.3 \mu\text{M}$ ), as does isocitrate lyase and PEP mutase<sup>17</sup>.

3,3-Difluorooxaloacetate, which differs from oxaloacetate by replacement of the C(3) hydrogens with fluorine atoms, is not a substrate for OAH. However, this substance is an exceptionally tight binding competitive inhibitor ( $K_i = 68 \pm 4 \text{ nM}$ ) (Figure 4). In contrast, malate is a weak binding competitive inhibitor of OAH. Malate differs from the oxaloacetate in that its C-2 center is tetrahedral and functionalized with a hydroxyl group. (*R*)-Malate ( $K_i = 2.5 \pm 0.2 \text{ mM}$ ) binds an order of magnitude tighter than does its *S*-enantiomer ( $K_i = 22 \pm 3 \text{ mM}$ ), but ~5-orders of magnitude less tightly than does 3,3-difluorooxaloacetate.

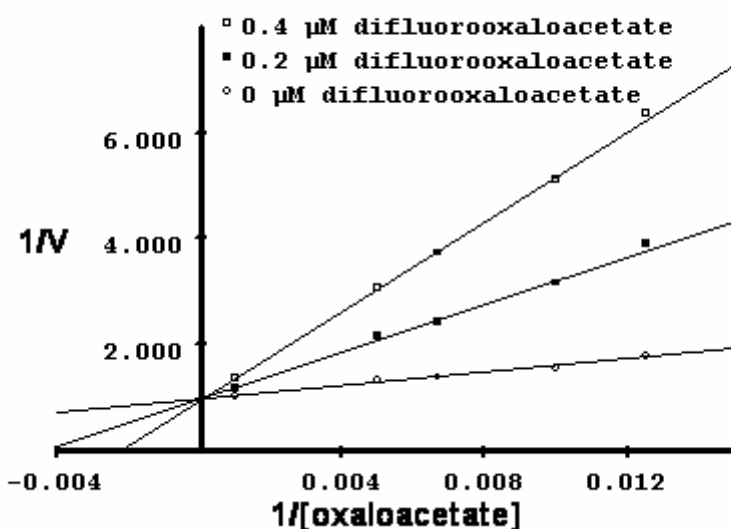


Fig. 4 The double reciprocal plot of the initial velocity of OAH catalyzed hydrolytic cleavage of oxaloacetate vs the concentration of oxaloacetate at changing, fixed concentrations of difluorooxaloacetate. See Materials and Methods for details.

## 5 Discussion

### 5.1 OAH Function and Distribution

The observations described above show that (1) oxalate formation in the fungi *A. niger* and *B. cinerea* derives from OAH catalyzed hydrolytic cleavage of oxaloacetate, and (2) the respective genomes of *A. niger* and *B. cinerea* contain a single OAH encoding gene. Although, the two OAHs share 70 % sequence identity, they display different behavior. The *A. niger* OAH is reported to associate to form high order oligomers and to be specific for  $Mn^{+2}$  as the metal ion cofactor. Both enzymes do, however conserve the active serine residue (Ser260 of *B.cinerea*) known to contribute significantly to OAH catalytic efficiency<sup>27</sup>.

BLAST searches, in which the *A. niger* OAH sequence is used as query, have led to the identification of one or more close homologs (as defined by > 50% sequence identity) encoded in the *A. niger* genome, and in the genomes of other *Aspergillus* strains. The homologs of *A. niger* do not possess the active site serine found in authentic OAH. Also, the genetic data presented above show that these proteins do not contribute to oxalate formation. Presently, their catalytic functions are unknown.

*A. niger* and *B. cinerea* are representatives of different subphyla of *ascomycete* fungi, separated by hundreds of millions of years of evolution. Our results strongly suggest that both fungi use OAH catalyzed hydrolytic cleavage of oxaloacetate as the main (if not the sole) route for oxalate production. Among the 22 fungi, whose genomes have been sequenced, the presence of the *oah* gene is strictly

correlated with oxalate production<sup>27</sup>. The acquisition of OAH activity in a given fungal species for oxalate production appears to be important for niche adaptation.

We searched for evidence of the occurrence of OAH in organisms from other kingdoms. Earlier reports indicate that oxalate biosynthesis occurs in some plants, especially in plants of the genus *oxalis*<sup>17,28-31</sup>. Studies carried out with plant tissue extracts indicate that enzyme catalyzed conversion of oxaloacetate to oxalate does take place (30). However, owing to the scarcity of sequenced plant genomes, plant *oah* genes have not yet been identified. One exception is the gene encoding the petal death protein of the flowering plant *Dianthus caryophyllus* (Swiss-Prot entry Q05957). The petal death protein possesses OAH activity in addition to 2-alkylmalate C-C bond lyase activity<sup>17</sup>. A BLAST search of the gene data bank, identified homologs of ~60% sequence identity in the plants *Arabidopsis thaliana* and *Oryza sativa*. However, neither of these homologs have been isolated or tested for activity and neither appear to possess an active site Ser which is well correlated with the presence of OAH activity<sup>27</sup>.

The presence of OAH activity has been observed in extracts of the bacterium *Streptomyces cattleya*<sup>32</sup>. Unfortunately, the sequence of the gene associated with this activity has not been reported. To identify other bacterial proteins that might possess OAH activity, we carried out a BLAST search of bacterial genomes using the *D. caryophyllus* petal death protein and the *B. cinerea* OAH as query. The *Bacillus cereus* Swiss-Prot entry Q738L6 was found to have the highest homology. This protein is similar in size (viz 302 amino acids) to the petal death protein and the *B. cinerea* OAH. Furthermore, it shares 39 % sequence identity with the petal death protein and 36% identity with the *B. cinerea* OAH. However, it does not appear that the *B. cereus* protein contains the conserved active site serine associated with OAH activity<sup>27</sup>. Rather, this protein has a proline residue at this position as do the other bacterial homologs of >35% sequence identity reported to date. The recombinant *B. cereus* Q738L6 was prepared by gene cloning and expression in *E. coli* and the purified enzyme was subjected to substrate screening<sup>33</sup>. The *B. cereus* Q738L6 was found to specifically catalyze the cleavage of 2-methylisocitrate to form succinate and pyruvate ( $k_{\text{cat}}/K_m = 2.5 \times 10^6 \text{ M}^{-1} \text{ s}^{-1}$ )<sup>33</sup>. Thus, the bacterial homolog is not OAH but instead 2-methylisocitrate lyase of the propionate pathway.

Based on the results of this survey we conclude that OAH is widely used among fungi for the purpose of oxalate formation associated with niche adaptation. Moreover, OAH may also function in oxalate producing plants and perhaps in some specialized bacteria.

## 5.2 OAH Catalysis and Inhibition

OAH catalyzes the hydrolytic cleavage of oxaloacetate (a retro-Claisen reaction) with high efficiency ( $k_{\text{cat}}/K_m = 1.5 \times 10^5 \text{ M}^{-1} \text{ s}^{-1}$ ) (Table 2). In addition, it promotes retro-Aldol reactions of 2R,3S-dimethylmalate and 2R-ethyl-3S-methylmalate with a 1000-fold lower efficiency ( $k_{\text{cat}}/K_m = 2 \times 10^2 \text{ M}^{-1} \text{ s}^{-1}$ ) (Table 3). The plant petal death protein displays these same types of catalytic activities but it promotes both processes with equal catalytic efficiency:  $k_{\text{cat}}/K_m = 2 \times 10^4 \text{ M}^{-1} \text{ s}^{-1}$  for hydrolytic cleavage of oxaloacetate and for lysis of 2R-ethyl-3S-methylmalate<sup>17</sup>.

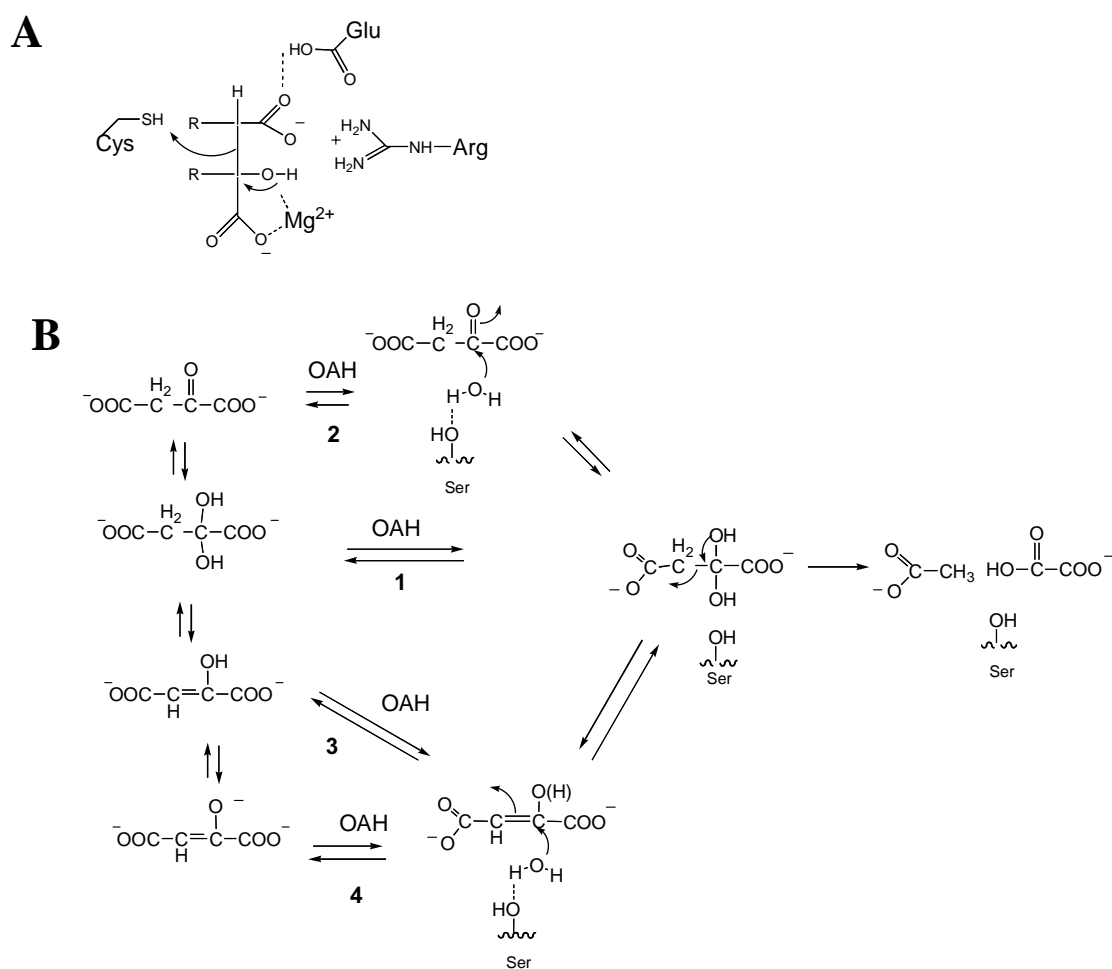
OAH and the petal death protein are members of the PEP mutase/isocitrate lyase enzyme superfamily<sup>17,34-40</sup>. Together with isocitrate lyase and 2-methylisocitrate lyase, they form a subgroup of enzymes of similar chemical function. Within this subgroup, the superfamily catalytic scaffold is tailored to bind and activate  $\alpha$ -oxocarboxylate metabolites for C $\alpha$ -C $\beta$  bond cleavage (17, 35-39). The ability of OAH and the petal death protein to act on C( $\alpha$ )OH and C( $\alpha$ )=O substrates is quite remarkable. This dual catalytic activity might also be possible for isocitrate lyase, but to our knowledge this has not been tested.

The catalytic mechanism that has been proposed for the isocitrate and 2-methylisocitrate lyases is depicted in Figure 5<sup>39</sup>. The key elements of catalysis are (1) electron withdrawal from the C( $\alpha$ )oxygen via electrostatic interaction with the  $\text{Mg}^{+2}$  cofactor and the conserved Arg residue, (2) electron withdrawal from the aci-carboxylate leaving group via hydrogen bond formation with the conserved glutamic acid residue, (3) general acid catalysis by the conserved Cys residue, and (4) general base catalysis by a yet to be assigned residue.

These elements are conserved in the petal death protein for which the active site structure is known<sup>17b</sup>. Although the X-ray structure of OAH is not yet available, it is evident from sequence alignment analysis that these same elements contribute to the OAH catalysis. Thus, the catalytic mechanism proposed for the isocitrate and 2-methylisocitrate lyases could be operative in the petal death protein and OAH mediated cleavage reactions of the 2R-alkylmalates. It is tempting to assume that the OAH catalytic scaffold is designed to promote reaction by way of a single catalytic mechanism (Figure 5, pathway 1). If so, the only adjustment required for operation of a retro-Aldol type mechanism would be the hydration of oxaloacetate to generate the C( $\alpha$ ) gem-diol. The C( $\alpha$ ) gem-diol is expected to bind in the active site in the same manner as do 2R-alkylmalates with the pro-S OH of the gem-diol located in place of the C( $\alpha$ ) alkyl groups of the malates. Modeling studies indicate that Ser257 in the petal death protein is properly positioned for interaction with the pro-S OH of the oxaloacetate hydrate<sup>27</sup>. The corresponding residue in OAH is Ser260.

Although the economy of this hydrate/retro-Aldol type mechanism is noteworthy, proof of its operation is needed. In theory, the OAH catalyzed oxaloacetate cleavage reaction might proceed by any one of several reasonable chemical pathways (Figure 5, pathways 1-4), each of which is

distinguished by the form of the oxaloacetate that serves as substrate. In aqueous solution at neutral pH, the keto-form of oxaloacetate (81 %) exists in equilibrium with the hydrate form (*i.e.*, the gem diol) (7 %) and the enol form (12 %) <sup>41</sup>. Oxaloacetate binds  $Mg^{+2}$  (and  $Mn^{+2}$ ) <sup>42</sup>. Consequently the ratio of the structural forms at equilibrium is changed, and owing to the reduction in  $pK_a$ , the enolate form may also contribute to this equilibrium <sup>42</sup>. At the current time, it is not clear whether the “free” enzyme binds the  $Mg^{+2}$ (oxaloacetate) complex or the enzyme- $Mg^{+2}$  complex binds the “free” oxaloacetate. Furthermore, the structural form of the oxaloacetate (ketone, gem-diol, enol or enolate) that is bound in the enzyme active site is not known. Therefore, each of the 4 chemical pathways depicted in Figure 5B must be considered.



**Fig. 5 A.** The catalytic mechanism of the  $\alpha$ -hydroxycarboxylate lyases of the PEP mutase/isocitrate lyase enzyme superfamily. **B.** The four pathways leading from oxaloacetate to oxalate and acetate.

Whereas the gem-diol is a minor component of solvated oxaloacetate, it is the major component by far of solvated 3,3-difluorooxaloacetate. In the latter case, only the gem-diol is observed by the  $^{13}\text{C}$ -NMR. However, the report that aspartate transaminase slowly catalyzes the conversion of difluorooxaloacetate to difluoroaspartate<sup>43</sup> suggests that the ketone form is present, if only in a trace amount. The high affinity exhibited by OAH for the competitive inhibitor 3,3-difluorooxaloacetate ( $K_i = 68\text{ nM}$ ; Figure 4) suggests that the enzyme preferentially binds the gem-diol form of oxaloacetate. Importantly, the gem-diol of oxaloacetate serves as a substrate in pathway 1 (Figure 5B), and as an intermediate in pathways 2 and 3. Only in pathway 4, involving the oxaloacetate enolate, does the gem-diol not serve as an intermediate. Interestingly, the electrostatic forces that are needed to stabilize the oxyanion intermediate formed in pathway 4, are not required to bind the gem-diol intermediate in pathways 1-3. The results of earlier isotope labeling studies, carried out with the OAH from *A. niger*, provide evidence that catalysis by this enzyme does not take place through the enol (or enolate). Thus, it appears that pathways 3 and 4 are not followed in the catalytic process.

Based on the fact that pathway 1 is followed in catalysis of the  $\text{C}(\alpha)\text{-C}(\beta)$  bond cleavage reactions of 2R-alkylmalates and that pathway 2, requires that the enzyme binds and activates a water molecule for nucleophilic attack at the  $\text{C}(\alpha)=\text{O}$ , we propose that the gem-diol retro-Aldol pathway is the most reasonable mechanism for the OAH catalyzed transformation of oxaloacetate to pyruvate and acetate. However, detailed studies are required to provide definitive proof for this mechanism.

## 6 References

1. Dutton, M. V., and Evans, C. S. (1996) *Can. J. Microbiol.* 42, 881-895
2. Gadd, G. M. (1999) *Adv. Microb. Physiol.* 41, 47-92
3. Nakagawa, Y., Shimazu, K., Ebihara, M., and Nakagawa, K. (1999) *J. Infect. Chemother* 5, 97-100
4. Godoy G, Steadman JR, Dickman B, Dam R. (1990) *Physiol. Mol. Plant. Pathol.* 37, 179- 191
5. Guimaraes, R. L., and Stotz, H. U. (2004) *Plant Physiol.* 136, 3703-11
6. Kirkland BH, Eisa A, Keyhani NO. (2005) *J Med Entomol* 42, 346-51
7. Munir, E., Yoon, J. J., Tokimatsu, T., Hattori, T., and Shimada, M. (2001) *Proc. Natl. Acad. Sci. U. S. A.* 98, 11126-30
8. Maxwell, D. P., and Bateman, D. F. (1968) *Phytopathology* 58, 1635-1642
9. Balmforth, A. J., and Thomson, A. (1984) *Biochem. J.* 218, 113-118
10. Hammel, K. E., Mozuch, M. D., Jensen, K. A., Jr., and Kersten, P. J. (1994) *Biochemistry* 33, 13349-54
11. Lenz, H., Wunderwald, P., and Eggerer, H. (1976) *Eur. J. Biochem.* 65, 225-236
12. Kubicek, C. P., Schreierl-Kunar, G., Wohrer, W., and Rohr, M. (1988) *Appl. Environ. Microbiol.* 54, 633-637
13. Van den Hombergh, J. P., Van de Vondervoort, P. J., Van der Heijden, N. C., Visser, J. (1995) *Curr. Genet.* 28, 299-308
14. Ruijter, G. J., van de Vondervoort, P. J., Visser, J. (1999) *Microbiology* 145, 2569-76
15. Pedersen, H., Christensen, B., Hjort, C., and Nielsen, J. (2000) *Metab. Eng.* 2, 34-41
16. Pedersen, H., Hjort, C., and Nielsen, J. (2000) *Mol. Gen. Genet.* 263, 281-286
17. Lu, Z., Feng, X., Song, L., Han, Y., Kim, A., Herzberg, O., Woodson, W. R., Martin, B. M., Mariano, P. S., and Dunaway-Mariano, D. (2005) *Biochemistry* 44, 16365-16376
18. Saxty, B. A., Novelli, R., Dolle, R. E., Kruse, L. I., Reid, D. G., Camilleri, P., Wells, T. N. C. (1992) *Eur. J. Biochem.* 202, 889-896
19. Alberg, D. G., Lauhon, C. T., Nyfeler, R., Fassler, A., and Bartlett, P. A. (1992) *J. Am. Chem. Soc.* 114, 3535-3546
20. De Graaff, L., van den Broeck, H., Visser, J. (1992) *Curr. Genet.* 22, 21-27
21. Kusters-van Someren, M. A., Harmsen, J. A., Kester, H. C., Visser, J. (1991) *Curr. Genet.* 20, 293-299
22. Pontecorvo G., Roper J.A., Hemmons L.M., Macdonald K.D., Bufton A.W (1953) *Adv Genet.* 5, 141-238.
23. Kars, I., Wagemakers, C.A.M., McCalman, M., van Kan, J.A.L. (2005) *Mol. Plant Pathol.* 6, 641-652.
24. Erlick, H. A., Ed. (1992) *PCR Technology Principles and Applications for DNA Amplification*, W. H. Freeman and Co., New York
25. Appel, R. D., Bairoch, A., and Hochstasser, D. F. (1994) *Trends Biochem. Sci.* 19, 258-260
26. Homberg, U., Hoskins, S. G., Hildebrand, J. G. (1995) *Cell Tissue Res.* 279, 249-259
27. Joosten H.J., Han Y., Weiling Niu W., Juan Du J., Vervoort J., Dunaway-Mariano D, Schaap PJ. Submitted
28. Weir, T. L., Bais, H. P., Stull, V. J., Callaway, R. M., Thelen, G. C., Ridenour, W. M., Bhamidi, S., Stermitz, F. R., Vivanco, J. M. (2006) *Planta.* 223, 785-795
29. Guo, Z., Tan, H., Zhu, Z., Lu, S., Zhou, B. (1999) *Plant Physiol. Biochem.* 41, 47-92
30. Chang, C., and Beevers, H. (1968) *Plant Physiol.* 43, 1821-1828
31. Millerd, A., Morton, R. K., and Wells, R. E. (1963) *Biochem. J.* 88, 276-281
32. Houck, D. R., and Inamine, E. (1987) *Arch. Biochem. Biophys.* 259, 58-65
33. Han, Y. (2006) Doctor of Philosophy Dissertation, University of New Mexico, NM
34. Huang, K., Li, Z., Jia, Y., Dunaway-Mariano, D., and Herzberg, O. (1999) *Structure Fold Des* 7, 539-548.
35. Chaudhuri, B. N., Sawaya, M. R., Kim, C. Y., Waldo, G. S., Park, M. S., Terwilliger, T. C., and Yeates, T. O. (2003) *Structure (Camb)* 11, 753-64.
36. Britton, K., Langridge, S., Baker, P. J., Weeradechapon, K., Sedelnikova, S. E., De Lucas, J. R., Rice, D. W., and Turner, G. (2000) *Structure Fold Des* 8, 349-362.
37. Sharma, V., Sharma, S., Hoener zu Bentrup, K., McKinney, J. D., Russell, D. G., Jacobs, W. R., Jr., and Sacchettini, J. C. (2000) *Nat Struct Biol* 7, 663-668.
38. Grimm, C., Evers, A., Brock, M., Maerker, C., Klebe, G., Buckel, W., and Reuter, K. (2003) *J Mol Biol* 328, 609-621.

## Chapter 4

39. Liu, S., Lu, Z., Han, Y., Melamud, E., Dunaway-Mariano, D., and Herzberg, O. (2005) *Biochemistry* 44, 2949-2962.
40. Chen, C.H., Han Y., Niu, W, Kulakova, A. N., Howard, A., Quinn, J. P., Dunaway-Mariano, D., Herzberg, O. (2006) *Biochemistry* 45, 11491-11504.
41. Emly, M. and Leussing, D. L. (1981) *J. Amer. Chem. Soc.* 103, 628-634.
42. Mao. H.K. and Leussing, D.L. (1981) *Inorg. Chem.* 20, 4240-4247.
43. Briley, P.A., Eisenthal, R., Harrison, R., and Smith G. D. (1977) *Biochem. J.* 161, 383-387.
44. Lenz, H., Wunderwald, P., Eggerer, H. Eur (1976). *J. Biochem.* 65, 225-236.

### **Footnotes**

The abbreviations used are: OAH, oxalacetate acetylhydrolase; PEP, phosphoenolpyruvate; IPTG, isopropyl- $\beta$ -D-thiogalactopyranoside; SDS-PAGE, sodium dodecyl-sulfate-polyacrylamide gel electrophoresis PCR, polymerase chain reaction; K<sup>+</sup>Hepes, potassium salt of N-(2-hydroxyethyl)piperzine-N'-2-ethanesulfonic acid; LDH, lactate dehydrogenase; DTT, dithiothreitol; NADH, nicotinamide adenine dinucleotide; HPLC, high performance liquid chromatography.



## Chapter 5

Comulator: A new tool for calculating correlated mutations.

## **Comulator: A new tool for calculating correlated mutations.**

### **1. Abstract**

The vast number of protein sequences and protein structures that are currently available has led to the development of the automated procedures for generating structure based multiple sequence alignments of proteins that belong to the same superfamily. These automatically created alignments can be composed of thousands of sequences and contain a wealth of useful evolutionary fingerprints, such as correlated mutations. To detect correlated mutations in large superfamily alignments a robust algorithm called Comulator was developed that is insensitive to alignment imperfections that result from the automation. Here it is demonstrated that the algorithm can be used to detect amino acid networks that are important for protein function and that the algorithm can be used to predict the effects of ‘network’ mutations. The existence of such amino acid networks was validated by mutagenesis experiments of phospho-glucose isomerase from *Pyrococcus furiosus*.

### **2. Introduction**

Proteins evolve within a framework of functional constraints that limit amino acid substitution possibilities at individual alignment positions. 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 additional compensating mutations for specific residue changes. The concept of these, so called, correlated mutations (CM) is quite simple but, due to alignment imperfections for instance, the detection proves more difficult. Therefore, a number of different algorithms have been developed that are able to screen alignments for these CMs. An overview of these algorithms is described by. All these algorithms systematically compare two alignment positions and calculate a score (CM score) that expresses the correlation. The CM score depends on the quality of the alignment as well as on the phylogenetic distribution of the alignment sequences. Methods, such as the mutual information method<sup>2,3</sup>, statistical coupling analysis<sup>4,5</sup> or the perturbation method<sup>6</sup>, compare amino acid distributions to a random distribution model and therefore require phylogenetically equal distributed alignments. Asymmetric alignments that contain a large group of very similar sequences, for instance, will give inequitably high CM scores. Some algorithms, such as the Pearson correlation method<sup>7</sup>, therefore make use of alignment weight factors to reduce the influence of groups of very similar sequences. Inaccuracies in the alignment, however, will influence these CM scores.

The Comulator CM algorithm was originally developed as an extension of 3DM<sup>8</sup>. 3DM is a program that can automatically build a molecular class specific information system (MCSIS)<sup>9</sup>. Due to the exponential growth of sequence databases large numbers of (predicted) protein sequences are available for most superfamilies. 3DM can automatically build structure based superfamily alignments of these sequences. Due to 3DM's automation of sequence collection the alignment can be composed of a phylogenetically asymmetric set of sequences and due to the automation of the alignment procedure the final alignment can contain locally misaligned sequences. Comulator is a robust algorithm insensitive for these alignment imperfections. The reason is that Comulator ranks the CM scores calculated for all possible alignment position pairs instead of expressing CM scores as statistical probabilities. A CM score calculated by Comulator is a measure for the degree in which specific amino acids tend to appear together within two alignment positions. A high CM score is given to an alignment position pair(x,y) if specific amino acids at position y appear more frequent in a sub-alignment, composed of only sequences with a specific amino acids at position x, than in the full alignment. The distance between two observed frequencies will be influenced by the introduction of a large set of very similar- or mis-aligned sequences, but it is assumed that by ranking, the highest CM score will still reflect the most significant correlation and that these correlations are biologically meaningful if the alignment contains at least a number of phylogenetically distant sequences.

This work demonstrates how CM scores as calculated by Comulator can be used to assign function to amino acids, and to predict the effects of mutations, which is demonstrated by the introduction of specific mutations at these correlating amino acid positions. Comulator was already applied to two relatively small superfamily alignments, both containing a few hundred sequences. For the nuclear receptor (NR) superfamily high CM scores were assigned to residues involved in co-factor binding<sup>8</sup>. For the isocitrate lyase-like/phosphoenolpyruvate mutase superfamily high CM scores led to the discovery of sub-family specific residues important in substrate binding specificity. Here Comulator was applied to alignments of two larger superfamilies: the pectin lyase-like superfamily, which is dominated by hydrolases, and the RmlC-like cupin superfamily containing over 1200 and 2000 sequences, respectively. Comulator can freely be used at: <http://3dmcsis.systemsbiology.nl/> comulator

## 2. Methods

### 2.1 Alignment generation.

Superfamily alignments that were used as input for Comulador were generated by 3DM<sup>8</sup> using default parameters. Briefly, in the first step 3DM automatically generates a structure based alignment that contains only structurally conserved positions (core positions). This alignment is generated by superimposing proteins with known structure that have the overall fold typical for the superfamily. In the second step the primary sequences of these structures are used as templates to build separate subfamily alignments with (predicted) protein sequences for which no structure is available. 3DM finally builds a superfamily alignment that contains only core positions by combining the separate subfamily alignments using the alignment of superimposed structures as a scaffold. Next a general 3D-numbering scheme is applied to the core positions. Here the method was applied to the RmlC-like cupin superfamily and the pectin-lyase like superfamily (nomenclature according to the SCOP database<sup>10</sup>). Protein structures belonging to these superfamilies were collected using the SCOP database combined with blast searches with the PDB database. The proteins of the pectin-lyase like superfamily consist of large sheets of repetitive  $\beta$ -strands that fold into a right-handed parallel  $\beta$ -helix (Fig. 1). Due to the high repetitive nature of the structure of these proteins multiple highly similar superposition solutions exist and therefore the program did not correctly superimpose all structures (Fig. 1). A small N-terminal helix could be used as a reference to manually adjust incorrect superimposed structures with YASARA (<http://www.yasara.com>).



**Fig. 1 C-alphas of four pectin-degrading enzymes after automatic superposition.**

*Pectin degrading enzymes have a small N-terminal  $\beta$ -helix (here on top) followed by a large  $\beta$ -strand. The automatic superpositioning resulted into two distinct groups: the green and the yellow structure form one correctly superposed group and the red and the blue structure form a second group. These different groups are superposed incorrectly, because they are shifted one strand. This shift is revealed by a shift of the N-terminal  $\beta$ -helices.*

## 2.2 CM algorithm

Comulator's CM algorithm is an extension of a method described for the detection of allosteric interactions in the nuclear receptor superfamily<sup>11</sup>. The underlying method is known as the statistical coupling analysis method<sup>4,5</sup> that calculates a CM score for two positions ( $x, y$ ) by selecting the most prevalent residue at position  $y$  and defines a sub-alignment consisting of only sequences having that specific amino acid at position  $y$ . The ultimate  $CM_{(xy)}$  score is based on the difference in amino acid distributions between the full alignment and the subset.

Comulator calculates CM scores according to formula 1. For each alignment position comulator calculates the frequency of each individual amino acid. For each over represented amino acid (default setting: in more then 5% of the sequences) at position  $x$ , the algorithm loops over all different amino acid types present at position  $y$ . For each amino acid type at position  $y$  the algorithm then calculates the difference between two fractions ( $F$ )(formula 1).  $F$  is the difference between the fraction of an amino acid at position  $y$  in the complete alignment compared to the fraction of the same amino acid in a subset of the alignment. This subset is composed of all sequences that have a specific amino acid at position  $x$ . If  $F$  is positive for a particular amino acid pair at position  $x$  and  $y$ , these amino acids tend to be present together in these two alignment positions and are therefore correlated. If  $F$  is negative for an amino acid pair, this specific combination of amino acids tends to be absent at positions  $x$  and  $y$ . The absence of amino acid combinations also contributes to the CM score between the two alignment positions. Comulator calculates CM scores for all possible alignment position pairs. The alignment position pair with the highest CM score was set to 1 and all other CM scores expressed as a fraction of 1 and ranked accordingly. The resulting scores are visualized in heatmaps build as interactive HTML pages. YASARA was used to visualize CM scores in a 3D environment.

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

### Formula 1. Mathematical formula of Comulators CM scores algorithm.

$N_x$  is the observed frequency of sequences having a certain amino acid at position  $x$ .  $N_y$  is the observed frequency of sequences having a certain amino acid at position  $y$ . Together they form amino acid pair  $xy$ .  $N_{yx}$  is the observed frequency of sequences having this amino acid couple at positions  $xy$ .  $N_{(tot)}$  is the total number of sequences in the alignment.

( $N_x/N_{(tot)}$ ) is the fraction of an amino acid at position  $x$

( $N_{yx}/N_y$ ) is the fraction of an amino acid at position  $x$  in a subset of sequences having only a specific amino acid at position  $y$ .

CMA is a value for the degree of correlation between position  $x$  and  $y$ . CM is a summation of absolute fraction differences ( $|F|$ ) multiplied by the observed frequency of each amino acid pair ( $N_{yx}$ ).

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

The cloning of *pgiA* is described by Verhees *et al.*<sup>12</sup>. Mutants were generated with the QuikChange Site-Directed Mutagenesis Kit (Stratagene, USA) following the manufacturers instructions with the following adaptations: 25 PCR cycles were applied, and the PCR mixture was incubated with *DpnI* 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)

Mutation		QuikChange primers
3DM #	PfPGI #	
P27A	P132A	FW: (5'- GTAGTTTATGTTCCCGCCTATTGGGCTCATAGG -3') RV: (5'- CCTATGAGCCCAATAGGCGGGAACATAAACTAC -3')
Y28G	Y133G	FW: (5'- GTAGTTTATGTTCCCCCGGTTGGGCTCATAGGACGG - 3') RV: (5'- CCGTCCTATGAGCCCAACCGGGGGAACATAAACTAC -3')
P27A/Y28G	P132A/Y133G	FW: (5'- GTAGTTTATGTTCCCGCGGTTGGGCTCATAGGACGG -3') RV: (5'-CCGTCCTATGAGCCCAACCGGCGGGAACATAAACTAC -3')
P27E/Y28G	P132E/Y133G	FW: (5'- GTAGTTTATGTTCCCGAAGGTTGGGCTCATAGGACGG -3') RV: (5'- CCGTCCTATGAGCCCAACCTTCGGGAACATAAACTAC -5')
P27R/Y28G	P132R/Y133G	FW: (5'- GTAGTTTATGTTCCCGCGGTTGGGCTCATAGGACGG -3') RV: (5'- CCGTCCTATGAGCCCAACCGCGGGAACATAAACTAC -5')

**Table 1. Primers used for the mutagenesis studies of *pgiA*.** Both the 3DM alignment position numbering (1<sup>st</sup> column) and the number of the corresponding position in the ORF of phospho-glucose isomerase from *P. furiosus* (2<sup>nd</sup> column) are indicated.

*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<sub>600</sub> of 0.5. Isopropyl-  $\beta$ -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 at 6000 amplitude. The cell-free extract was clarified by centrifugation (37,000 g at 4 °C for 20 min). *E. coli* proteins were denatured by incubating the cell free 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 to 1 M NaCl in 20 mM Tris.HCl (pH 8.0). Fractions containing PGI were pooled, concentrated and loaded on a pre-equilibrated Superdex 200 GL column. Protein eluted in 20 mM Tris.HCl containing 125 mM NaCl.

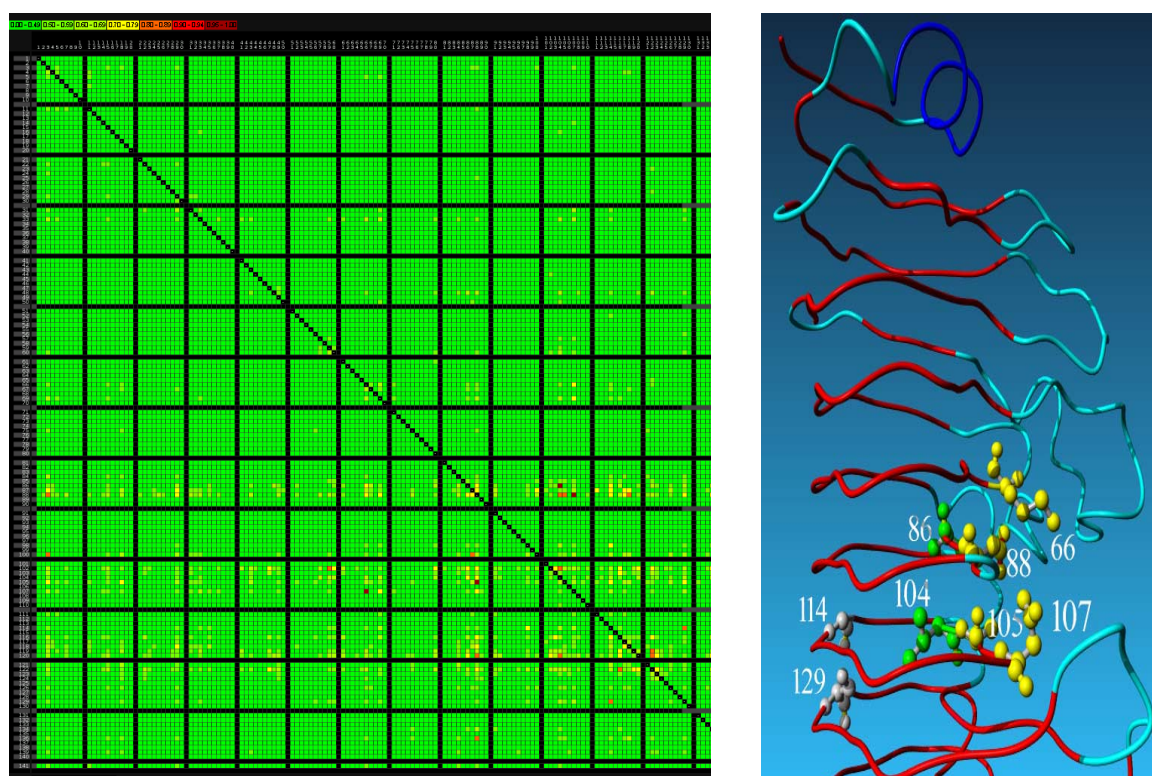
The determination of the enzyme activity of the different PGI mutants on fructose 6-phosphate was performed at 50°C as described previously<sup>12</sup> with the following adaptations: 20 mM Tris.HCl pH 7.0 was used, and the protein samples were incubated with 50mM EDTA at 50 °C for 20 minutes to

ensure complete metal depletion of the different samples<sup>22</sup>. Activity was measured with  $\text{MnCl}_2$  in excess over EDTA to ensure enzyme saturation.

### 3. Results and discussion

#### 3.1 *Pectin-lyase like superfamily*

Enzymes involved in the modification and cleavage of the pectin backbone are generally referred as pectinases. Although differing in their substrate specificity (homogalacturonan, xylogalacturonan, rhamnogalacturonan) and reaction mechanism (hydrolases, lyases, esterases), these proteins share the same single stranded right handed parallel  $\alpha$ -helix topology (Fig. 1, Fig. 2: right panel) characteristic for the pectin lyase-like superfamily. Notwithstanding the name a large portion of the collected pectin-lyase like protein sequences encode hydrolases (>65%: based on annotations). The pectin-lyase like proteins bind the polymeric substrate to an exterior cleft that starts after five consecutive turns of the right-handed helix. As with other depolymerizing enzymes, such as proteases and nucleases, the substrate binding cleft of these proteins consists of multiple sub-sites, each one of which makes contact with one monomeric unit of the polymeric substrate. The actual cleavage occurs between sub-site -1 and +1, while sub-sites, distant from the bond that undergoes catalysis, modulate the activity, affinity and/or specificity of the enzyme<sup>13</sup>. The most invariant position within the alignment is position 85 (3D-numbering scheme), which is an aspartate in more than 78% of the alignment sequences. It has been demonstrated that amino-acids at this position are directly involved in catalysis, such as the conserved catalytic aspartate in galacturonases (Asp180 in the protein sequence of PGII from *A. niger*)<sup>14,15</sup> or the lysine in pectate lyases (Lys190 in protein sequence of PelC from *E. chrisanthemi*)<sup>16</sup>. The substrate binding cleft of these proteins consists of different so called subsites on the outside of the proteins where the sugar residues of the pectin bind. This substrate binding cleft always starts after about five consecutive turns of the right-handed helix. Almost all correlations with CM scores above 0.8 are found in the substrate binding cleft. There are only two exceptions located within the N-terminal part outside the substrate binding cleft (Fig. 2: left panel).



**Fig. 2** Left panel: heat map of CM scores of the pectin-lyase like superfamily alignment. Alignment positions are indicated on the x- and y axes. The bar at the top indicates the coloring scheme (from green, representing low CM scores, to dark red, representing high CM scores). Right panel: Tube representation of endo-polygalacturonase II of *A. niger*. The N-terminal helix is represented in blue. The  $\beta$ -sheets that form the right-handed  $\beta$ -helix fold are red. The positions with high CM scores are represented in ball and stick. Note that from a visual scan of the left panel the positions in the substrate binding cleft can easily be spotted. In the projection of the structure in the right panel they are approximately on the same height.

The highest CM scores were detected for position 88. In polygalacturonases, alignment position 88 is often an aspartate (Asp183 in the protein sequence of *A. niger* PGII) that is involved in maintaining the proper ionization state of the catalytic aspartate at position 85. Mutation of this residue strongly decreases both the specific activity and the affinity of the enzyme to its substrate<sup>17</sup>. In pectin lyase A and B from *A. niger*, position 88 is occupied by a tyrosine (Tyr215 in the protein sequence of PLA from *A. niger*) that, together with several other Tyr and Trp residues, is involved in shielding the solvent and thereby increasing the basicity of the catalytic arginine at position 102 (Arg236 of in the protein sequence of PLA from *A. niger* numbering)<sup>18</sup>. It is striking that although glycoside hydrolases and pectin lyases adopted different catalytic mechanisms, position 88 still has the same function, e. a. decreasing the pKa of a catalytic residue to values corresponding to the optimal pH at which the reaction takes place.

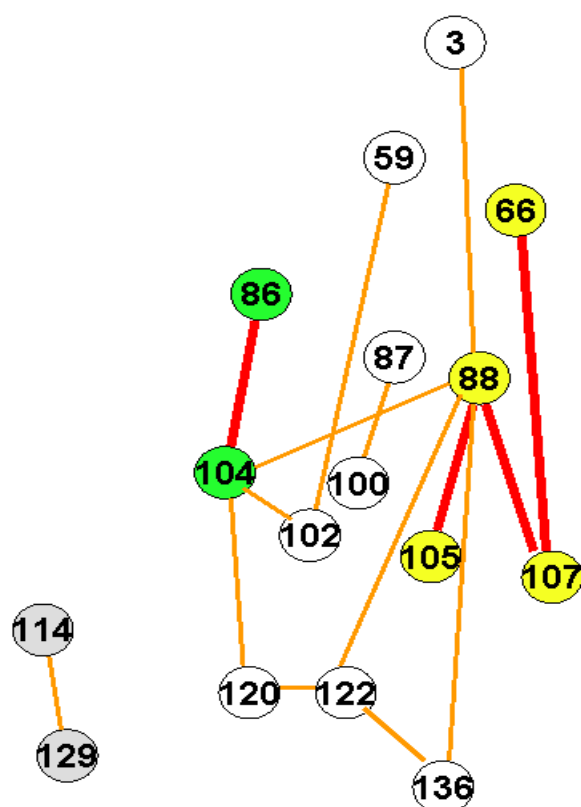
Alignment position 88 is the center of a network of high correlating positions (Fig. 3). Within this network position 66, 105 and 107 have CM scores above 0.95. Covering three consecutive turns of the right-handed helix of the overall protein fold, the side chains of the amino acids at these positions all point to the exterior of the protein near each other surrounding position 88 (Fig. 2: right panel, yellow



residues). The combination of high CM scores with the structural positional relation of these residue positions strongly suggests a functional relation.

Separately, a CM score above 0.95 was detected for alignment positions 86 and 104. These two positions are located facing each other on different  $\alpha$ -strands separated only one consecutive turn of the right-handed  $\alpha$ -helix. The side chains of the corresponding residues point to the interior of the right-handed  $\alpha$ -helix (Fig 2: green residues). These residues at these positions might serve a structural role, important for the proper orientation of the neighboring catalytic residues. Interestingly position 104 in pectin lyases (Pro238 in the protein sequence of *A. niger* PLB) and pectate lyases (Pro220 in the protein sequence of *E. chrisanthemi* PelC) corresponds to a proline with an unusual *cis* conformation<sup>19</sup>.

Alignment positions 114 and 129 form a separate network of two positions with CM scores above 0.8 (Fig. 3). These two positions are another example of residues that face each other from opposite  $\alpha$ -strands separated by 1 turn of the overall right-handed  $\alpha$ -helix fold (Fig. 2: grey residues). This position pair is only very loosely connected to the previously described network through positions 114 and 107 with a weak CM score of 0.63.



**Fig. 3. Network of all alignment positions with correlations above 0.8.** Nodes represent alignment positions. Note that their position (N-terminus on top) and their colors (high correlating positions ( $>0.95$ ) in yellow and green, the separate network formed by 114 and 129 in grey) are similar to the representation of the protein in figure 2. The network edges are colored according their CM score (see Fig. 2 panel A for the scale). Note that only two positions (3 and 59) with a CM scores above 0.8 were detected in the N-terminus outside the substrate binding cleft.

Using annotations extracted from the Swiss-Prot database we were able to link specific amino acid contents of correlating positions with specific protein functions (Table 2). There are 5 functional groups. Motif 66[Q,W]86[G]88[D]104[I,V]105[A]107[K,N]

is an evolutionary fingerprint that applies to most of the hydrolases of this superfamily while motif, 66[F]86[G]88[H]104[I,V]105[S]107[G] applies for a subset of plant hydrolases. Motif 66[G]86[V]88[L]104[P]105[R]107[R] applies for lyases and motif 66[D]86[F]88[F]104[T]105[A]107[G] for esterases. Although no experimental data could be found in the literature that proves the role of most of the discussed positions, these examples show how the algorithm can detect functionally important residues, which can be used to guide protein engineering experiments.

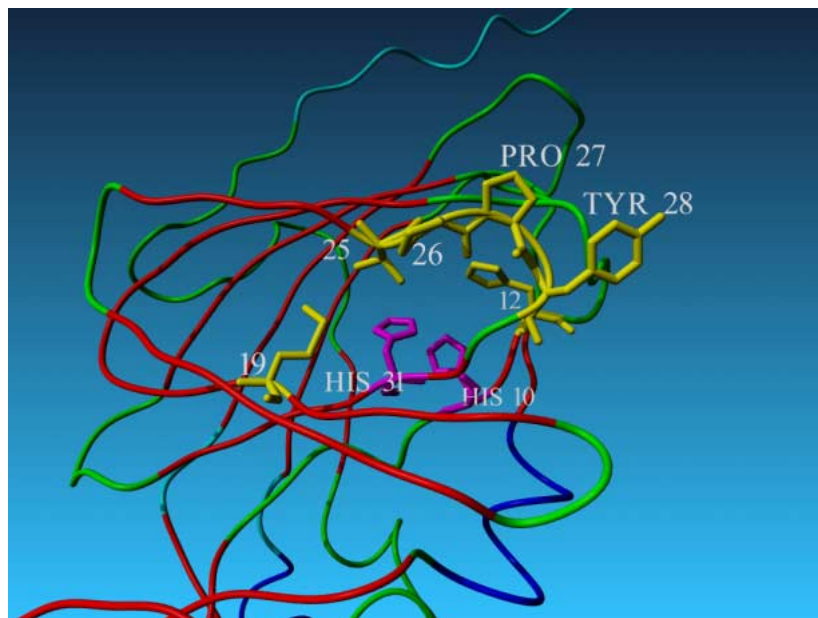
Amino acid pair (3D-numbers)	% sequences with specific AA pair	Most abundant reaction type (Swiss-Prot annotation)	% of sequences with specific reaction type
Positions 86 and 104			
G86 I104	24 %	Hydrolases	98 %
G86 V104	26 %	Hydrolases	96 %
V86 P104	12 %	Lyases	85 %
F86 T104	7 %	Esterases	100%
Positions 88 and 107			
D88 K107	18 %	Hydrolases	95 %
D88 N107	17 %	Hydrolases	98 %
H88 G107	12 %	Hydrolases (plant)	87 %
L88 R107	15 %	Lyases	82 %
F88 G107	6%	Esterases	95 %
Positions 88 and 105			
D88 A105	32%	Hydrolases	98 %
H88 S105	19%	Hydrolases (plant)	92 %
L88 R105	11%	Lyases	80 %
F88 A105	7%	Esterases	100%
Positions 66 and 107			
F66 G107	11%	Hydrolases (plant)	96 %
Q66 N107	13%	Hydrolases	97 %
W66 K107	11%	Hydrolases	95 %
G66 R107	13%	Lyases	79 %
D66 G107	6%	Esterases	100%

**Table 2: Relation between the most abundant amino acid types at high correlating positions and the function of proteins.** Highly correlating position pairs are indicated in the grey bars. For each pair the most left two columns lists the most abundant amino acid pairs together with the percentage of proteins in the alignment that have this specific amino acid couple. This enzyme superfamily can be divided in four major classes of enzymatic activities. The third column displays the most abundant class for each amino acid couple according to Swiss-Prot annotations. The last column displays the percentage of proteins with this annotation and the specific amino acid couple.

Most of the sequences of the pectin-lyase like protein superfamily are annotated as hydrolases (>60%). The remaining sequences are classified as lyases (~22%), esterases (~13%) and others (~4%). To test the effect of a large set of similar sequences on CM scores we performed CMA analyses on reduced sets of input sequences. In an iterative process we randomly removed hydrolases from the alignment. Up to 80% of the hydrolases could be removed without having a significant effect on the overall composition of the CM networks.

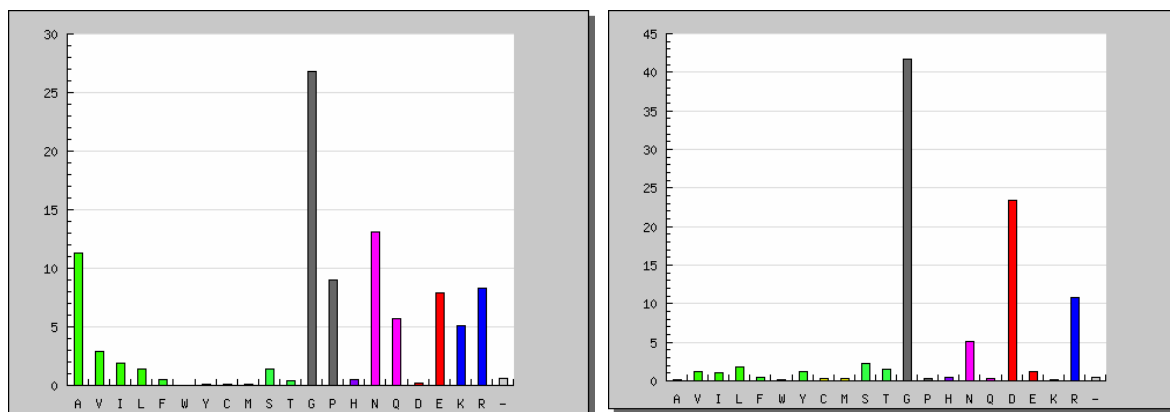
### 3.2 RmlC-like cupin superfamily

The alignment of the cupin superfamily is the largest of the four superfamilies and contains 2097 different sequences. The RmlC-like cupin superfamily consists of proteins possessing a common  $\alpha$ -barrel structure also known as a jelly roll (cupin) fold. Although the proteins in this superfamily are functionally diverse<sup>20</sup>, most proteins in this superfamily are enzymes of which the active site is located within the barrel of the proteins. This active site often contains two histidines (3D-numbers 10 and 31) that are conserved in approximately 80% of all sequences (Fig. 4: magenta residues). CMA analyses of the cupin superfamily alignment revealed a network of highly correlating positions (CM scores > 0.95) composed by alignment positions with 3D-numbers 12, 19, 25, 26, 27 and 28. In most members of the superfamily alignment positions 25-28 form a structurally conserved loop (loop25-28) located at the outside of the protein, in the direct neighborhood of histidine 31 (Fig 4: yellow residues).



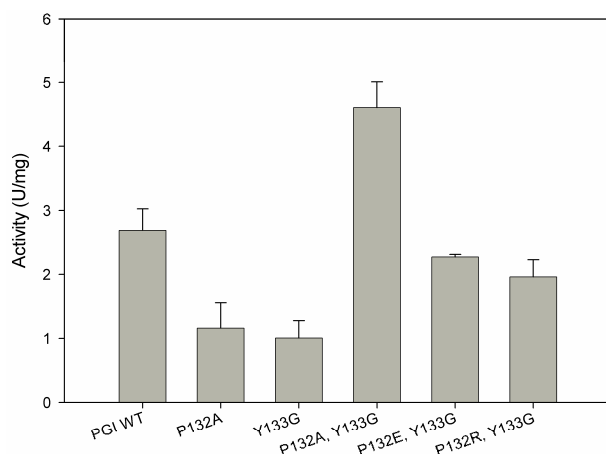
**Fig 4. Tube representation of the 3D-structure of PfPGI from *P. furiosus* (PDB accession code: 1X82).** Helices are blue,  $\alpha$ -strands are red, loops green, the two conserved histidines are magenta and core positions with high CM scores are yellow.

Mutational data is available for the cupin superfamily. However the network of correlated positions has barely been touched. Mutagenesis of alignment position 28 in flavonol synthase from *Citrus unshiu* (G261A) resulted in 95% reduction of enzyme activity. Introduction of a proline resulted in a completely non active enzyme<sup>21</sup>. One of the best characterized members of the cupin superfamily is the *Pyrococcus furiosus* PGI (PfPGI)<sup>22-24</sup>. Several crystal structures of this protein have been elucidated<sup>22-25</sup> and the reaction mechanism has been analyzed by mutagenesis, NMR and EPR studies<sup>24</sup>. PfPGI has a tyrosine at alignment position 28 (Fig. 4). Glycine is the most prevalent amino acid at alignment position 28 within the cupin super family alignment (42%, see Fig. 5). Interestingly, mutagenesis of this tyrosine to a glycine in PfPGI (Y133G) results in a 2.3 fold reduction of the activity (Fig. 6). The highest CM score was detected for alignment position pair 27 and 28. The most abundant residue at position 27 is the glycine (present in 26% of all sequences). In PfPGI it is a proline residue. Although a proline is present in only 8% of all superfamily sequences, 66% of sequences with a tyrosine at alignment position 28 have a proline at position 27. Although the combination G27G28 (3D-numbers) is the most expected amino acid combination based on the abundance of the amino acids at these two positions, G27G28 is present in only 1% of all sequences. In fact, an alanine is the most abundant residue (24%) in sequences with a glycine at position 28. The combination A27G28 is present in 10% of all superfamily sequences.



**Fig. 5.** Bar graphs representing the amino acid distributions of positions 27 (left panel) and 28 (right panel). On the x-axes are the 20 different amino acids and on the y-axes the percentage of sequences that have a specific amino acid.

It was therefore hypothesized that the introduction of an alanine at position 27 might compensate for the negative effect of mutation Y28G. Although single mutations P27A and Y28G (in PfPGI: P132A and Y133G) both have a negative effect on the activity of the PfPGI the double mutant A27G28 resulted in a protein with an almost doubled activity compared to the wild type. The second most abundant residue in sequences with a glycine at position 28 is the glutamate (18.5%) closely followed by the arginine (18.2%). Again, both double mutants E27G28 and R27G28 regained activity (Fig. 6).



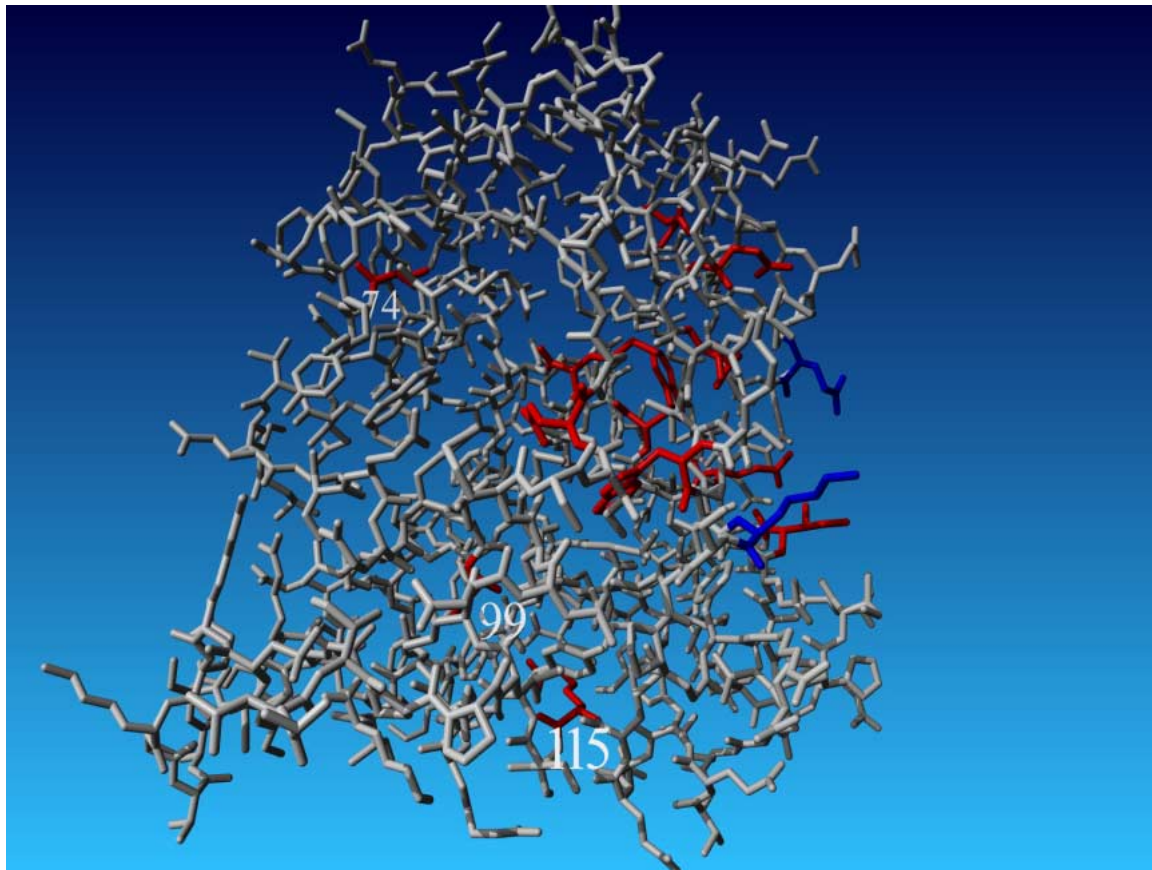
**Fig. 6. Bar graphs representing activity of single or double mutants of wild type PfPGI.** Numbering according to the amino acid sequence of PfPGI.

### 3.3 Isocitrate lyase-like/Phosphoenolpyruvate mutase superfamily.

The Isocitrate lyase/PEP mutase superfamily alignment contains a number of different enzymes that catalyze widely varying types of reactions. What they have in common is that they all act on a carbon and an oxalate-like moiety. Structurally the enzymes in this superfamily share an  $\alpha$ -barrel overall fold. The structure based multiple sequence alignment created by 3DM contains 170 structurally conserved core positions and is composed of 375 unique sequences. Nine residue positions were detected that had CM scores above 0.8. These residue positions are found mainly but not exclusively surrounding the active site. The 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 one subfamily, the oxaloacetate hydrolase (OAH) subfamily. Mutating this serine (3D-number 157) to a threonine or proline (the most prevalent residues in other sequences in the alignment) did not significantly decrease  $V_{max}$ , but had drastic effects on the affinity of OAH for oxaloacetate<sup>27</sup>. This serine is required for OAH activity and could be used as a marker to distinguish between OAH orthologs and closely related paralogs<sup>27</sup>.

### 3.4 Nuclear receptor superfamily.

Nuclear receptors are transcription factors that are activated by agonistic ligands (hormones). This hormonal activation induces a conformational change, which enables the binding of a protein (co-factor) after which the transcription starts. The NR superfamily alignment contains 750 unique sequences. CMA analyses revealed a network of twelve residue positions (3DM alignment numbers 14, 26, 30, 33, 39, 55, 59, 62, 66, 74, 99, 115) that showed CM scores higher than 0.8 with at least 2 other positions<sup>8</sup>. Structurally 9 of these 12 positions are located near the co-factor binding site (Fig. 7). All these 9 positions show decreased ability of co-factor binding upon mutating<sup>8</sup>. Though the other 3 positions are not located at the co-factor binding site, mutational analyses of two of these positions (99, 115 (Fig. 7)) appeared to have a major effect on co-factor binding capacity<sup>8</sup>. For the last position (74) no evidence for a role in co-factor binding is available.



**Fig. 7 Representation of highly correlating positions in the NR ligand-binding domain.**

*Residues with high CMA scores are in red. Note that most of these residues (except for residues with alignment number 74, 99, 115) cluster near the cofactor-binding site here marked by the conserved lysine of helix 3 (in blue) and the conserved glutamate of helix 12 (in blue). These residues are believed to form a charge clamp to which the co-factor can bind.*

### 3.5 WWW version

The WWW based version of Comulator accepts as input aligned fasta and ClustalW format files. The input alignments are visualized similarly as for 3DM derived alignments including an alignment

positions numbering scheme. CM scores are visualized as an interactive heatmap (Fig. 2: left panel). If a sequence contain a Swiss-Prot protein ID they are automatically linked to the corresponding Swiss-Prot data file. In addition all Swiss-Prot amino acid features are collected from the Swiss-Prot database and a link is created from the specific amino acids to the corresponding Swiss-Prot data file. All alignments and CM heat maps described in this work can be downloaded at the 3DM website ([www.3dmcsis.systemsbiology.nl](http://www.3dmcsis.systemsbiology.nl)).

### **4. Conclusions**

Comulator is a fast, robust and powerful tool for the detection of correlated mutations in huge automatically created superfamily alignments. Comulator is insensitive to alignment imperfections that are a result of automation of alignment generation. Comulator was tested on four superfamily alignments consisting of 300 - 2100 protein sequences. Site directed mutagenesis experiments revealed that Comulator could be used to predict residues important for substrate specificity, residues involved in the binding of a co-factor and compensating mutations.

## 5. 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. Clarke ND. Covariation of residues in the homeodomain sequence family. *Protein Sci*. 1995;4:2269-2278.
3. Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW. Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis. *Mol Biol Evol*. 2000;17:164-178.
4. Lockless SW, Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*. 1999;286:295-259.
5. Fodor AA, Aldrich RW. On evolutionary conservation of thermodynamic coupling in proteins. *J Biol Chem*. 2004;279:19046-19050.
6. Dekker JP, Fodor A, Aldrich RW, Yellen G. A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments. *Bioinformatics*. 2004;20:1565-1572.
7. Gobel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. *Proteins*. 1994;18:309-317.
8. Joosten HJ, Folkertsma S, Zimmeren F, Lutje Hulsik DJ, Kuipers R, Ittmann E, Roijen E, Vriend G, Schaap PJ. 3DM: A new generation of molecular-class-specific information systems applied to four protein super-families
9. Horn F, Vriend G, Cohen FE. Collecting and harvesting biological data: the GPCRDB and NucleaRDB information systems. *Nucleic Acids Res* 2001;29:346-349.
10. SCOP: a Structural Classification of Proteins database. Hubbard TJ, Ailey B, Brenner SE, Murzin AG, Chothia C. *Nucleic Acids Res*. 1999;27:254-256.
11. Shulman AI, Larson C, Mangelsdorf DJ, Ranganathan R. Structural determinants of allosteric ligand activation in RXR heterodimers. *Cell* 2004;116:417-429.
12. 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 superfamily. *J Biol Chem*. 2001 Nov 2;276(44):40926-32. Epub 2001 Aug 30.
13. Davies GJ, Wilson KS, Henrissat B. Nomenclature for sugar-binding subsites in glycosyl hydrolases. *Biochemical Journal* 1997;321:557-559.
14. van Santen Y, Benen JA, Schroter KH, Kalk KH, Armand S, Visser J, Dijkstra BW. 1.68-Å crystal structure of endopolygalacturonase II from *Aspergillus niger* and identification of active site residues by site-directed mutagenesis. *J Biol Chem* 1999;274:30474-30480.
15. Armand S, Wagemaker MJ, Sanchez-Torres P, Kester HC, van Santen Y, Dijkstra BW, Visser J, Benen JA. The active site topology of *Aspergillus niger* endopolygalacturonase II as studied by site-directed mutagenesis. *J Biol Chem* 2000;275:691-696.
16. Scavetta RD, Herron SR, Hotchkiss AT, Kita N, Keen NT, Benen JA, Kester HC, Visser J, Jurnak F. Structure of a plant cell wall fragment complexed to pectate lyase C. *Plant Cell* 1999;11:1081-1092.
17. Pages S, Heijne WH, Kester HC, Visser J, Benen JA. Subsite mapping of *Aspergillus niger* endopolygalacturonase II by site-directed mutagenesis. *J Biol Chem* 2000;275:29348-29353.
18. Sanchez-Torres P, Visser J, Benen JA. Identification of amino acid residues critical for catalysis and stability in *Aspergillus niger* family 1 pectin lyase A. *Biochem J* 2003;370:331-337
19. Vitali J, Schick B, Kester HCM, Visser J, Jurnak F. The Three-Dimensional Structure of *Aspergillus niger* Pectin Lyase B at 1.7-Å Resolution. *Plant Physiol*. 1998;116:69-80.
20. Dunwell JM, Purvis A, Khuri S. Cupins: the most functionally diverse protein superfamily. *Phytochemistry* 2004;65:7-17.
21. 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.
22. 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 Aug 29;278(35):33290-7. Epub 2003 Jun 9.
23. Berrisford JM, Akerboom J, Brouns S, Sedelnikova SE, Turnbull AP, van der Oost J, Salmon L, Hardre 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 Oct 22;343(3):649-57.
24. 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 May 19;358(5):1353-66. Epub 2006 Mar 24.
25. Hansen T, Oehlmann M, Schonheit P. Novel type of glucose-6-phosphate isomerase in the hyperthermophilic archaeon *Pyrococcus furiosus*. *J Bacteriol.* 2001 Jun;183(11):3428-35.
  26. Akerboom J, Turnbull AP, Hargreaves D, Fisher M, de Geus D, Sedelnikova SE, Berrisford JM, Baker PJ, Verhees CH, van der Oost J, Rice DW. Purification, crystallization and preliminary crystallographic analysis of phosphoglucose isomerase from the hyperthermophilic archaeon *Pyrococcus furiosus*. *Acta Crystallogr D Biol Crystallogr.* 2003 Oct;59(Pt 10):1822-3. Epub 2003 Sep 19.
  27. 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. (Submitted)

## 6. Appendix

Appendix 1: List of PDB files used for the creation of the superfamily alignments.

SCOP superfamily classification	PDB accession numbers from structure files used for superpositioning and determination of core positions	files used as template for 3DMSA creation
Pectin-lyase like superfamily	1AIR,1BHE,1BN8,1CZF,1DAB,1DBG,1DBO,1EE6,1GQ8,1H6U,1H80,1HG8,1IA5,1IB4,1IDJ,1IDK,1JRG,1JTA,1K5C,1KCC,1KCD,1NHC,1O88,1O8D,1O8H,1OFL,1OFM,1OOC,1PCL,1PE9,1PLU,1PXZ,1QCX,1QJV,1RMG,1TSP,1VBL,1XG2,2BSP,2EWE,2PEC,	1BHE,1DBO,1GQ8,1H80,1HG8,1IB4,1JRG,1NHC,1PXZ,1QCX,1QJV,1RMG,1TSP,1XG2,2BSP,2EWE
RmlC-like cupin	1BK0,1BLZ,1CAU,1CAV,1CAW,1CAX,1DCS,1DGR,1DGW,1DRT,1DRY,1DS0,1DS1,1DZR,1DZT,1E5H,1E5I,1E5S,1EP0,1EPZ,1EYB,1FI2,1FXZ,1GP4,1GP5,1GP6,1GQG,1GQH,1GQW,1GVG,1GY9,1H1I,1H1M,1HB1,1HB2,1HB3,1HB4,1HJF,1HJG,1IPJ,1IPK,1IPS,1J1L,1J3P,1J3Q,1J3R,1J58,1JR7,1JUH,1L3J,1LKN,1LR5,1LRH,1M4O,1NXM,1NYW,1NZC,1O5U,1OBN,1OC1,1OD5,1ODM,1ODN,1OFN,1OI6,1OIH,1OII,1OIJ,1OIK,1OS7,1OTJ,1PHS,1PM7,1PML,1QIQ,1QJE,1QJF,1QXJ,1QXR,1QY4,1RTV,1RXF,1RXG,1UCX,1UD1,1UIJ,1UIK,1UNB,1UO9,1UOB,1UOF,1UOG,1UPL,1UW8,1UZW,1VR3,1VZ4,1VZ5,1W03,1W04,1W05,1W06,1W28,1W2A,1W2N,1W2O,1W3V,1W3X,1W9Y,1WA6,1WLT,1X7N,1X82,1X8E,1XJA,1Y3T,1ZRR,2AAC,2ARA,2ARC,2BU9,2CAU,2CAV,2ET1,2ET7,2ETE,2PHL,	1DGR,1DRT,1DZT,1EPZ,1EYB,1FXZ,1GP6,1GQG,1GQW,1HJG,1IPJ,1J3Q,1J58,1JR7,1LRH,1NXM,1OD5,1OFN,1OIJ,1PM7,1QJE,1RTV,1UNB,1VR3,1WA6,1WLT,1X82,2ET7

## Summary

In the industry where proteins are utilized in a wide range of different applications these proteins normally need optimization. Rational *in silico* protein engineering is a relatively inexpensive tool that can guide the optimization of these proteins. Moreover, such a tool can also be applied in drug discovery processes where the understanding of the interactions between amino acids of a target protein with potential drugs is crucial. However rational *in silico* protein engineering has proven to be a great challenge. Integrated protein engineering platforms that can effectively be used to design mutation strategies that lead to improved protein functionality are still scarce. This thesis describes the development of 3DM, a prototype rational *in silico* protein engineering platform designed to guide scientist in protein engineering studies.

Superfamily data contains a wealth of useful information that can guide protein engineering. The amount of data available of one superfamily is nowadays enormous and the collection, storage, ordering, updating and analyzing this data can be a huge task. 3DM is a program that can do the most laborious steps automatically and combines existing analysis methods in one protein engineering platform. The key feature of this system is a relational database based upon the Molecular Class Specific Information System (MCSIS) technology invented in 1992. Superfamily databases built by 3DM revolve around a structure based superfamily alignment that is used for the transfer of amino acid related data, such as mutational information, from well-studied protein members to other proteins of the superfamily. Moreover, the setup of the 3DM databases makes the alignment data also applicable for detection of evolutionary traces, such as correlated mutations. In this thesis 3DM was applied to a number of different superfamilies each with a specific research question.

**Chapter 2** describes 3DM applied to four superfamilies. These four examples clearly show the power of having all superfamily data in a structured database. A 3DM MCSIS database was built for the P53 tumor repressor protein superfamily for which we were able to collect almost 30.000 mutations. The type of cancer is known for almost all these mutations and therefore it was possible to link residue conservation to disease. For some alignment positions, for instance, specific amino acid mutations that led to cancer were amino acid residues normally not observed at the corresponding alignment position. Moreover, it was shown that from the 30.000 mutations, 25.000 were found in structurally conserved (or core) regions of the protein. This is on average of 161 mutations per core position. In comparison, in the structurally variable positions this was only 22. Another 3DM database was built for the

nuclear receptor (NR) superfamily. This is a good example how superfamily alignments can be used for the transfer of mutational information from one sequence to the other. In contrast to the P53 superfamily, where all recorded mutations originate from human P53 protein, here mutations have been introduced in multiple protein members. *Ab initio* correlated mutation analyses predicted a network of 12 protein positions important in co-factor binding. For 10 of these 12 alignment positions we were able to automatically collect validating mutational information from the literature. This clearly shows that correlating amino acid positions hint at functional relationships and that these properties of alignments can be used to predict amino acid function. To prove this, we also mutated one of the amino acid positions not yet covered by the scientific literature. A predicted amino acid change at this position completely disrupted co-factor binding.

**Chapter 3 and 4** present detailed work on the characterization of oxaloacetate hydrolase (OAH). OAH catalyzes the conversion of oxaloacetate into oxalate and acetate. By gene-disruption and complementation experiments combined with comparative genomics we show that this conversion is the main route used by fungi to produce oxalate. OAH is a member of the phosphoenolpyruvate/pyruvate superfamily. Analysis of the 3DM superfamily alignment led to the discovery of an OAH specific serine residue essential for catalysis, which was verified by mutagenesis experiments. In addition it was shown that this serine residue can serve as a marker to distinguish OAH from paralogous sequentially very similar OAH class proteins.

**Chapter 5** describes a new correlated mutation analysis scoring algorithm specifically developed for the analyses of automatically created large superfamily alignments and its application on two new superfamilies. Here it was demonstrated that this algorithm can deal with phylogenetically skewed alignments and can bypass local alignment imperfections. It was also demonstrated that the algorithm can be used to identify networks of functionally related amino acids. Validation of these networks came from literature studies and by mutagenesis experiments.

**Final remark:** this thesis gives an indication of what will be possible in the near future in the field of rational protein engineering. The work on 3DM is far from finished. 3DM can potentially be applied to hundreds of superfamilies, harboring proteins and enzymes that are important for many different applications. Realizing all its unused potential, hopefully this thesis marks only the beginning of the development of the 3DM rational protein engineering platform.

## Dankwoord

Na vier jaar kan ik met zekerheid zeggen dat promoveren alles behalve eenvoudig is. Het vereist dusdanig veel doorzettingsvermogen, creativiteit en expertise dat ik dit werk nooit alleen had kunnen doen. Op de lange weg die uiteindelijk geleid heeft tot het schrijven van dit dankwoord hebben velen mij geholpen en ik wil dan ook iedereen hartelijk bedanken. Een aantal mensen wil ik graag bij name noemen.

Laat ik beginnen met mijn begeleider Peter Schaap. Peter, karakteristiek aan jou is jouw directe “Amsterdamse” aanpak. Ik mij kan voorstellen dat sommige mensen hier aan moeten wennen, maar ik denk (en jij zult het hier wel mee eens zijn) dat het tussen ons vanaf het begin van mijn promoveren goed is verlopen. Natuurlijk hebben wij wel eens de nodige aanvaringen gehad: Ik heb moeten leren om nee te zeggen tegen de oneindige stroom van nieuwe ideeën waarmee jij komt aandraven. Toch kan ik over het algemeen terug kijken op een uiterst fijne samenwerking gedurende de afgelopen vier jaar dat met succes wordt afgerond in de vorm van dit proefschrift. Ik hoop dat dit geen eindpunt is, maar een opstapje.

Als tweede wil ik mijn promotor Prof. Johan van den Berg bedanken. Johan, jij bent erg behulpzaam geweest. 3DM wordt vercommercialiseerd. Je bent altijd erg enthousiast geweest en ook zeer betrokken bij de opstartfase van het bedrijf. Wij hebben samen de “business challenge” gewonnen ondanks dat we ons niet echt goed hebben kunnen voorbereiden. We hebben samen heel Nederland doorkruist en hebben hier veel goede connecties aan overgehouden. Ook heb je veel tijd gestoken in het patenteren van 3DM en je hebt een potentieel eerste klant (BIOVET) weten te interesseren en hopelijk zullen zij binnenkort met ons in zee gaan. Ik wil je voor je inzet hartelijk bedanken.

Prof. Gert Vriend. Gert, jij verdient misschien nog wel het grootste deel van dit dankwoord. Vooral omdat jij vaak voor mij klaar hebt gestaan en dit altijd op vrijwillige basis hebt gedaan. Ik heb gebruik mogen maken van jouw enorme kennis en gedrevenheid voor de wetenschap. We hebben uren samen aan artikelen geschreven en aan het einde van mijn onderzoek heb je veel tijd gestoken in het schrijven van een STW voorstel voor een vervolg traject van dit onderzoek. Ik hoop van harte dat dit voorstel gehonoreerd zal worden en we nog lang kunnen werken aan dit onderwerp. Ik hoop ook dat je nauw betrokken wilt blijven bij het vercommercialiseren van 3DM, want ik denk dat de kans op succes vele malen groter is met jouw naam er aan verbonden.

Aan de ontwikkeling van 3DM en labwerk dat verricht is aan OAH hebben acht studenten hard meegewerkt. Zonder hun was dit werk zeker nooit tot stand gekomen. Dit zijn Judith, Debora, Frank, Miao Miao, Remco, Erik0, Erik1, Manickam en Imre. Super bedankt voor jullie inzet en het vele werk dat jullie samen hebben verzet. Ik denk dat we niet alleen veel bereikt hebben maar ook een gezellige tijd hebben gehad. Vooral in het poolcentrum, of niet Erik? We zullen het nog eens overdoen.

Dan zijn er natuurlijk nog mijn collega's die altijd klaar hebben gestaan als ik iets nodig had. Ook zij hebben mij een goede tijd bezorgd. Ik zal de "vrijdagmiddag" discussies met Harry en Elena niet vergeten. Zeker niet omdat die de hele week konden plaatsvinden. Ik ben blij dat ik jullie heb leren kennen en denk dat onze trip naar Japan een goede impuls is geweest voor onderlinge relaties. Dit geldt ook voor Jorg. Helaas heb jij Jorg, vroegtijdig gekozen om ergens anders te gaan werken. Ik heb een leuke tijd met jou gehad als "labtafelgenoten".

Kelly, jij hebt het altijd voor elkaar gekregen om mij de hele weg naar huis te laten watertanden met je gesprekken over eten. Ik vond het altijd erg gezellig om met jou, Serve en Bart tussen Nijmegen en Wageningen heen en weer te rijden. Je zult tegen de tijd dat je dit leest al lang bevallen zijn. Ik wens jullie heel veel geluk met de nieuwe aanwinst.

Douwe, ik vind jou echt een prachtkerel met je scherpe tong en een goed gevoel voor humor. Jij wordt/bent absoluut een goede wetenschapper. Daar mag je nooit aan twijfelen.

Deze mensen waren natuurlijk niet de enige die tegelijk met mij hebben gewerkt bij fungal genomics. Ik wil ook de rest van de "fungenners" bedanken voor een fijne tijd en alle geboden hulp.

Tijdens dit project is veel samen gewerkt met andere groepen zoals biochemie waar vooral Jacques Vervoort en Marc van den Kamp erg behulpzaam waren. Jacques, bedankt voor het kritisch lezen van dit manuscript. Binnen microbiologie hebben we nog samengewerkt met de groep van John van der Oost. John, ook jij bedankt voor de samenwerking en het kritisch lezen van dit manuscript. Jasper, Ik denk dat jouw mutanten één van de beste resultaten van dit proefschrift zijn. Bedankt hiervoor. Vanuit de universiteit van New Mexico hebben Prof. Debra Dunaway-Mariano en Dr. Ying Han een grote contributie geleverd aan het natte werk dat is verricht aan expressie en mutagenese van oxaloacetate hydrolase.

Ik ben erg trots dat Simon en Olaf toegezegd hebben mijn paranimfen te zullen zijn. We zijn niet alleen verbonden door de wetenschap, maar bovenal goede vrienden. Bedankt jongens.

Naast iedereen die in werksfeer betrokken was met dit onderzoek is er natuurlijk mijn lieve familie (papa, mama, Mariska en Reinout) die ik niet mag vergeten. Jullie hebben altijd in mij geloofd. Ik wil jullie hiervoor hartelijk bedanken.

Als laatste mijn grote liefde Repke. Groot respect gaat van mij uit hoe jij onze Yara op de wereld hebt gezet. BEDANKT:

Tussen allen en alleen,  
één en geen,  
maar één  
wereld groot verschil.  
Wij twee  
Ons drie

## Curriculum Vitae

Personalia	
Name	Hendrik Johannes Joosten
Address	Weezenhof 8064
Zip-code and place	6536 CP Nijmegen
Telephone (private)	06-53548336
Telephone (work)	0317-486013
Date of birth	23-08-1975
Place of birth	Amersfoort
Nationality	Dutch
Sofi-nr.	193128986
Email address	henk-jan.joosten@wur.nl

Education	
1994-1999	Hoger Laboratorium Onderwijs (HLO), organic chemistry
2000-2003	Katholieke Universiteit Nijmegen (KUN) faculty of Molecular Life Science (MLW), bioinformatics.
2003-2007	A.I.O, micro-biology, university of Wageningen

Work experience	
1998-1999	My practical training period as organic chemist (HLO) was performed at the University of Nijmegen, where I worked on the synthesis of several enantiomeric pure bis-oxazolines copper complexes, which are catalysts for asymmetric cyclopropanation of styrenes
1999-2001	I worked as organic chemist at MercaChem (Nijmegen). I worked on the syntheses of a large variety of potential new drugs for several companies. I worked on multiple step syntheses, determination of the best route for synthesis and small scale combinatorial chemistry.
2001-2003	During my practical training as bio-informatition at the CMBI (University of Nijmegen) I was part of the MCSIS (Molecular Class Specific Information system) development team. Within this project I developed automated homology modeling program and worked on the development of the NuclearDB (an MCSIS for the nuclear receptor superfamily) in collaboration with Organon. During this project I worked on multiple sequence alignment techniques, 3 dimensional superposition of protein structures, programming in python and turbo-pascal, the usage of a PostgreSQL database and other relevant bioinformatic-programs (e.g. WHATIF, Yasara).
2003-2007	During my PhD at fungal genomics (microbiology, Wageningen University) I mainly worked at the development of 3DM: A protein engineering platform. 3DM was used in several protein engineering projects. For one project I did the labwork myself, for the other 3DM projects labwork was performed via several collaborations.



## List of publications

### Full papers:

1. P.Smeets, W. Fleuren, H.J. Joosten. Transport of para-aminohippurate, fluorescein, and fluoresceinmethorexate by multidrug resistance proteins 2 and 4. *Netherlands Journal of Drug Research*, 2002, 5, 3-6.
2. Folkertsma S, van Noort P, Van Durme J, Joosten HJ, Bettler E, Fleuren W, Oliveira L, Horn F, de Vlieg J, Vriend G. A family-based approach reveals the function of residues in the nuclear receptor ligand-binding domain. *J Mol Biol.* 2004 Aug 6;341(2):321-35.
3. Han Y, Joosten HJ, Niu W, Zhao Z, Mariano PS, McCalman MT, van Kan JA, Schaap PJ, Dunaway-Mariano D. Oxaloacetate hydrolase: The C-C bond lyase of oxalate secreting fungi. *J Biol Chem.* 2007 Jan 23 [Epub]
4. Joosten HJ, Folkertsma S, Zimmeren van F, Lutje Hulsik DJ, Kuipers R, Ittmann E, Roijen E, Vriend G, Schaap PJ. 3DM: A new generation of molecular-class-specific information systems applied to four protein super-families. *PEDS* (Submitted)
5. 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* (Accepted)
6. Joosten HJ, Akkerboom J, Muthuraman M, Kuipers R, Oost van der J, Schaap PJ. Comulator: A new tool for calculating correlated mutations. (In progress).

### Patents:

1. Joosten HJ, Kuipers R, Ittmann E, Roijen E, Schaap PJ. Europa octrooiaanvraag 06120830.2
2. Joosten HJ, Kuipers R, Ittmann E, Roijen E, Schaap PJ. Method of generating a protein database. Provisional Amerikaanse octrooiaanvraag 60/845,170

## Education/Conferences

	<i>ECTS</i>
<i>credits</i>	
<b>Courses</b>	
Radiation expert 5B, biochemie, Wageningen, 2003	1.4
Bioinformatics and drug design, Universiteit Nijmegen, 2003	2.6
Bioinformation Technology , VLAG, 2003	2.8
Advanced Microarray, UMC, 2005	1.5
Business challenge, WBG, Wageningen, 2005	1.5
Oriëntatie op ondernemerschap, STW, 2006	1.2
<b>Meetings</b>	
NBC, Rehorst, Ede, 2004	0.5
Exploring the edges of omics, EMBL, Germany, 2004	1.0
Systems biology, Heidelberg, Germany, 2004	1.0
EBC12, Copenhagen, Denmark, 2005	1.0
BBC, Gent, Belgium, 2005	1.0
Systems biology, Amsterdam, 2006	0.5
Protein Design and Evolution for Biocatalysis, Greifswald, Germany, 2006	0.5
Protein Engineering XIX, 2007, Vancouver, Canada	1.0
<b>General courses</b>	
Scientific writing, Wageningen University, 2005	1.7
<b>Optionals</b>	
Preparation PhD research proposal	6.0
Literature study program biochemistry group, 2001-2005, Wageningen	1.0
Bio-informatics platform Wageningen, 2005-2007	1.0
Preparation STW proposal	1.0
Fungal genomics meetings, Wageningen, 2003-2007	4.0
PhD/Postdoc meetings Laboratory of Microbiology, 2003-2007	2.0
Japan PhD trip, 2004	3.0
Meeting University of New Mexico, New Mexico, USA, 2006	1.5
<b>Total</b>	<b>37,7</b>

### **Financial support**

Research presented in this PhD dissertation was financially supported by The Graduate School VLAG (Voeding, Levensmiddelentechnologie, Agrotechnologie en Gezondheid)

The printing of this manuscript was sponsored by:

- Dr. Ir v.d. Laarstichting  
Postbus 4446  
6401 CX Heerlen
- Genencor International, B.V  
Archimedesweg 30  
2333 CN Leiden
- Wageningen Universiteit en Researchcentrum  
Wageningen Universiteit  
Postbus 9101  
6700 HB Wageningen

### **Printing**

Propress b.v.  
Vadaring 122  
6702 EB Wageningen

### **Cover design**

Erik Roijen