Automatic and manual functional annotation in a distributed web service environment

Inaugural-Dissertation

zur
Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln
vorgelegt von

Anika Jöcker

aus Haan

2009

Berichterstatter/in: Prof. Dr. Thomas Wiehe
Prof. Dr. Martin Hofmann-Apitius

Tag der letzten mündlichen Prüfung: 24 April 2009

Zusammenfassung

Während die Anzahl öffentlich verfügbarer genomischer Sequenzen stetig steigt, sind die meisten Gene nicht ausreichend funktionell charakterisiert. Die Bestimmung der Genfunktion und die Entdeckung funktionaler Beziehungen zwischen Genen wird die nächste große Herausforderung im post-genomischen Zeitalter.

In diesem Kontext sind einerseits verbesserte Pipelines und Programme notwendig, denn die Durchführung von Experimenten würde zu viel Zeit in Anspruch nehmen. Andererseits müssen automatische Vorhersagen manuell überprüft werden, um ihre Glaubwürdigkeit beurteilen zu können und um ein umfassenderes Bild über die Funktion jedes einzelnen Gens zu bekommen.

Häufig findet die automatische funktionale Annotation von Genen durch den Transfer von Funktionen von bereits funktional charakterisierten Genen statt, wobei Programme wie Blast benutzt werden. Allerdings hat dieser Ansatz viele Nachteile und macht systematische Fehler, da Speziations- und Duplikationsereignisse nicht mitberücksichtigt werden. Der phylogenomische Ansatz allerdings ist in der Lage die Vorhersagegenauigkeit wesentlich zu verbessern, indem die evolutionäre Geschichte von Genen mit in Betracht gezogen wird.

In dieser Arbeit wird der manuelle Prozess von der Assemblierung der DNS bis zu der funktionalen Charakterisierung von Genen und der Identifikation und dem Vergleich von syntänischen Regionen am Beispiel einer Region im Kartoffelchromosom V erklärt und Probleme diskutiert. Weiterhin werden Kandidatengene in der Region ermittelt, die bei der Pathogenresistenz eine Rolle spielen. Um die automatische funktionale Annotation in Genomprojekten zu verbessern, wird eine phylogenomische Pipeline vorgestellt, welche SIFTER, eins der besten phylogenomischen Programme, beinhaltet. Diese Pipeline wird verbessert und an den Genomen von *Medicago truncatula*, *Sorghum bicolor* und *Solanum lycopersicum* getestet. Um neue Kandidatengene herauszufinden, die zur Entwicklung von Medikamenten und Pflanzenschutzmitteln verwendet werden könnten, werden nicht-pflanzenspezifische Gene, wie zum Beispiel die Transferrin Familie, die bis jetzt in Pflanzen unbekannt war, aus dem Genom von *M. truncatula* und *S. bicolor* herausgefiltert und näher untersucht.

Um die Annotation weiter zu verbessern, wird ein neuer phylogenomischer Ansatz entwickelt. Dieser benutzt annotierte Funktionsattribute wie zum Beispiel Interaktionspartner, Proteindomänen usw., um die Funktionsmutationsrate zwischen Genen und Gengruppen in einem phylogenetischen Baum zu ermitteln und um herauszufinden, ob die Funktion von einem Gen oder einer Gengruppe auf ein anderes oder eine andere übertragen werden kann. Dieser neue Ansatz wird in das SIFTER Programm integriert und wird an der Blue-light photoreceptor/Photolyase Familie und an einem Testdatensatz von manuell kurierten *Arabidopsis thaliana* Genen getestet. Die Vorhersagegenauigkeit konnte für beide Datensätze signifikant verbessert werden.

Da Genfunktionen mit bioinformatischen Methoden nie mit hundertprozentiger Genauigkeit vorhergesagt werden können, wird das AFAWE System zur manuellen Annotation vorgestellt. In AFAWE werden verschiedene Web Services zur funktionalen Annotation gestartet und die Ergebnisse und Zwischenergebnisse so dargestellt, dass sie einfach zu vergleichen sind. AFAWE kann für jeden Organismus und jede Art von Gen verwendet werden. Aufgrund seiner flexiblen Struktur, können neue Web Services und Workflows leicht in AFAWE integriert werden. Zur Zeit ist neben Blast-Suchen in verschiedene Datenbanken und Programmen zur Suche von Proteindomänen, auch die phylogenomische Pipeline in AFAWE als Analyse verfügbar. Verschiedene Filter helfen dem Benutzer glaubwürdige Vorhersagen von unglaubwürdigen zu unterscheiden. Weiterhin kann eine detaillierte manuelle Annotation zu jedem Gen angegeben werden, welche dazu benutzt werden soll, die automatische Annotation in öffentlichen Sequenzdatenbanken wie MIPSPlantsDB zu ersetzen.

Abstract

While the number of genomic sequences becoming available is increasing exponentially, most genes are not functionally well characterized. Finding out more about the function of a gene and about functional relationships between genes will be the next big bottleneck in the post-genomic era.

On the one hand improved pipelines and tools are needed in this context, because running experiments for all predicted genes is not feasible. On the other hand manual curation of the automatic predictions is necessary to judge the reliability of the automatic annotation and to get a more comprehensive view on the function of each individual gene.

For the automatic functional annotation often a homology based function transfer from functionally characterized genes is applied using methods like Blast. However, this approach has many drawbacks and makes systematic errors by not taking care of speciation and duplication events. Phylogenomics has shown to improve the functional prediction accuracy by taking the evolutionary history of genes in a phylogenetic tree context into account.

In this thesis the manual process from the assembly of the DNA sequence to the functional characterization of genes and the identification and comparison of shared syntenic regions, including the identification of candidate genes for pathogen resistance in potato chromosome V, is explained and problems discussed. To improve the automatic functional annotation in genome projects, a phylogenomic pipeline, which includes SIFTER one of the best phylogenomic tools in this area, is introduced, improved and tested in the *Medicago truncatula*, *Sorghum bicolor* and *Solanum lycopersicum* genome projects. To obtain new candidate genes for the development of new drugs and crop protection products, non-plant specific genes, like the transferrin family which is not known in plants yet, are extracted from the *M. truncatula* and *S. bicolor* genomes and further investigated.

For further improvement of the annotation, a new phylogenomic approach is developed. This approach makes use of annotated functional attributes to calculate the functional mutation rate between genes and groups of genes in a phylogenetic tree and to find out if the function of a gene can be transferred or not. The new approach is integrated into the SIFTER tool and tested on the blue-light photoreceptor/photolyase family and on a test set of manually curated *Arabidopsis thaliana* genes. Using both test sets the prediction accuracy could be significantly improved and a more comprehensive view on the gene function could be obtained.

But because still no tool is able to annotate all functions of a gene with 100% accuracy, I introduce a system for manual functional annotation, called AFAWE. AFAWE runs different web services for the functional annotation and displays the results and intermediate results in a comprehensive web interface that facilitates comparison. It can be used for any organism and any kind of gene. The inputs are the amino acid sequence and the corresponding organism. Because of its flexible structure, new web services and workflows can be easily integrated. Besides Blast searches against different databases and protein domain prediction tools, AFAWE also includes the phylogenomic pipeline. Different filters help to identify trustworthy results from each analysis. Furthermore a detailed manual annotation can be assigned to each protein, which will be used to update the functional annotation in public databases like MIPSPlantsDB.

Acknowledgements

I am very grateful to many people for helping me attain my Ph.D. Without their help it would not have been possible for me to succeed.

First of all I like to express special thanks to Dr. Heiko Schoof who gave me the possibility to realize my ideas as a member of his bioinformatics research group at the Max-Planck Institute for Plant Breeding Research in Cologne. I am very grateful for his guidance, all the inspiring discussions, his support and his critical reading. From him I learned to work independently and I had the possibility to work together with many different people within the institute and around the world in research projects.

I also like to say special thanks to Prof. Dr. Thomas Wiehe from the Genetics Institute at the University of Cologne for being my supervisor, for his guidance, his support and the valuable discussions we had. I also like to thank him for all the commentaries and suggestions he gave me on my thesis and for his constitutive words.

My gratitude goes also to my co-advisors, Prof. Dr. Martin Hofmann-Apitius from the Fraunhofer Institute SCAI and Dr. Ulrike Göbel for their interest in the projects, their support and their critical reading of the thesis.

Thanks also to Prof. Dr. Martin Hülskamp for being the chairman of the examination committee.

Many thanks to Ph.D. Barbara Engelhardt and Prof. Ph.D. Steven Brenner for providing me with the SIFTER code and for very helpful discussions and advice. Special thanks also to Dr. Michael Plümer for helping me in questions concerning the SIFTER framework and Jens Warfsmann for discussions and reading.

Also I like to thank Dr. Agim Ballvora, members of the MIPS institute in Munich, members of the international tomato annotation group (ITAG) and members of the international Medicago annotation group (IMGAG) for successful collaboration. In particular Dr. Chris Town, Manuel Spannagl, Dr. Klaus Mayer and Dr. Remy Bruggmann I am deeply grateful for providing me necessary data and analysis results, for their support and their very helpful advice.

I am also indebted to my parents and my best friend Sabrina Heinhaus for their support. They were always willing to listen to my problems although they do not really understand what I am doing here.

Especially I like to thank my beloved husband Andreas, my workmate, flat mate, travel mate and constituent partner in many discussions. Thank you for your love, patience, support and motivation.

I am deeply grateful for all your comments and suggestions being the first reader of the chapters of my thesis. I dedicate this thesis to you. Thank you for the wonderful two years we had together in Cologne. I am looking forward to our future in Heidelberg and all the things we have planned.

Table of Contents

Tables

Figures

I. Introduction & Motivation

Since the first complete genomic sequence of an Eukaryote was published in 1996 [Goffeau et al. 1996], more and more sequence data is becoming publicly available and this data will increase in the next years, because next-sequencing technologies enable a fast and cheap sequencing of whole genomes [Hall 2007].

However, deciphering the DNA is just the first step in understanding the molecular machinery of an organism. Next steps include the detection and definition of gene structures and regulatory elements and their functional characterization. This information can afterwards be used to identify interactions between genes, to map genes to known regulatory and metabolic pathways or to search for candidate genes for further experiments.

The gene finding process in genome projects is done fully automatically by using bioinformatic algorithms which are trained for the considered organisms to increase the detection accuracy. In some projects the structural annotation is also checked and updated afterwards by a community to curate wrong predictions and to provide a useful dataset for further analyses [Thibaud-Nissen et al. 2007]. However, this is often not done in case of the functional characterization of genes. The automatic functional characterization of predicted genes in genome projects is often done by annotation transfer, which confers functional annotations to a query sequence from a putative homologous gene, which is already functionally characterized. Unfortunately this general method has many drawbacks and often leads to wrong functional assignments, which then are propagated through public databases [Galperin and Koonin 1998] [Gilks et al. 2002].

In this thesis different approaches for the automatic transfer of functions between genes are introduced, tested and discussed. To improve the automatic function transfer in genome projects an existing approach is extended, tested and integrated in a flexible workflow. The workflow is afterwards integrated into a system, which facilitates a fast comparison between results from different automatic functional annotation programs and enables scientists to add a manual functional annotation to each gene. These manual annotations will be used in the future to update the automatic functional annotation in genome projects.

The thesis is divided into nine chapters. This Chapter (Chapter I) gives a general introduction to the topic, whereas Chapter II explains the goals of the thesis. Background information about the thesis topic and the current status and limitations of approaches for automatic function prediction are introduced in Chapter III. Furthermore web services and web service workflows are explained, which enable the integration of additional functional information about genes. Web services are used by tools described in the thesis to increase the flexibility and scalability of tools and workflows. The manual process from the determination of the DNA sequence to the functional characterization of genes and the identification and comparison of syntenic regions including the identification of candidate genes for pathogen resistance in potato chromosome V is described in Chapter IV. Chapter V introduces an automatic pipeline used for the automatic functional annotation in the *Medicago truncatula*, *Sorghum bicolor* and *Solanum lycopersium* (tomato) genome projects, because a manual functional annotation would not be feasible for whole genomes. To further improve the automatic function annotation in these and other genome projects a phylogenomic tool SIFTER-X is described and tested in chapter VI. However, to verify the automatically annotated functions (no tool is 100% accurate) and to provide a more comprehensive view of the function of each gene in a genome, a flexible system called AFAWE is introduced in chapter VII. AFAWE incorporates different analysis tools and enables users to add manual annotations to genes by providing an intuitive web interface and different filters to highlight trustworthy results. Furthermore AFAWE is connected to the public sequence database MIPSPlantsDB. In Chapter VIII results from the different approaches are discussed and an outlook how the manual and automatic annotation could be further improved is given in Chapter IX.

II. Aim of the thesis

One goal of the project is the improvement of functional annotation in genome projects. That is, sensitivity should be increased while at the same time decreasing the false discovery rate. Increasing the sensitivity means that more true functions are added to more genes. However, a low false discovery rate indicates that most assigned functions are true and only few are wrong.

Another goal of the project is to facilitate the manual functional annotation of unknown protein coding genes from arbitrary origins for every scientist. The user should be able to give as input an amino acid sequence and the corresponding organism name. Different analyses, which are selected by the user, are run and the results are shown in an intuitive and easily comparable way.

In the beginning of the project different analysis tools have been tested in collaborations with biologists of the Max-Planck Institute for Plant Breeding Research to find the best of them and to detect problems. To enable a fast access and an easy ex-changeability and scalability of the different programs, web services and web service workflows have been implemented and tested afterwards in the *Medicago truncatula*, *Sorghum bicolor* and *Solanum lycopersicum* genome projects. Because it was found that additional functional information (e.g. interaction partners, domain information) can further improve the automatic function prediction [Xiao and Pan 2005] [Hsing et al. 2008] [Zhao et al. 2008] [Mostafavi et al. 2008] one tool is extended by this kind of information and tested afterwards on a manually checked test set.

To achieve the second goal, all tested tools and workflows have been integrated in a system, which facilitates a fast comparison between the analysis results. The most trustworthy results of each analysis are highlighted by applying dynamic thresholds to the results, so that the user does not need any knowledge about the output score from the analysis and can concentrate on the comparison of the results. In this context a fast and easy integration of new tools or workflows into the system is required. Because of that all analysis tools are run as web services, because web services improve the scalability, accessibility, maintainability, efficiency and simplify the process. Furthermore each user is able to add his/her own manual annotation to each gene. This manual annotation will be used afterwards to update the automatic functional annotation in genome projects.

III. Background

1) Automatic functional prediction

With the increasing amount of genomic data becoming available the need of tools for automatic function prediction and data integration will be the next big challenge in biological science. Running experiments to find the function for a few genes takes month to years and therefore running experiments for all 20000 to 40000 genes of an organism is not feasible. In this case an automatic pipeline is needed. However the automatic function prediction of genes is not easy, although a huge amount of genes and their functions are conserved in all organisms. The challenge is to find out which of the genes share the same functions and which do not and which functions can be transferred from genes with known functions to genes with unknown functions. A search for homologous genes in sequence databases via Blast [Altschul et al. 1997] can lead to wrong predictions in cases such as duplication events, gene loss, domain shuffling or errors in databases [Gilks et al. 2002] [Galperin and Koonin1998]. If there is no high sequence identity between two genes, which share the same functions, the function prediction becomes very hard. In this case other functional information like expression patterns, interaction partners, structure prediction, search for protein domains or gene-neighborhood analysis can give clues to the true function of the gene. This information is also valuable for validation of orthologous[1] relationships [Xiao and Pan 2005] [Hsing et al. 2008] [Zhao et al. 2008] [Mostafavi, et al. 2008]. If no sequence similarity to functionally characterized genes can be found other methods can be used instead like gene fusion (Rosetta stone method), phylogenetic profiling, amino acid composition or critical residues detection (for review see [Friedberg et al. 2007] and [Lee et al. 2007]). However, each of these approaches has limitations and by using only one of them the result is often restricted to a specific class of proteins [Karimpour-Fard et al. 2008].

In the last few years many hybrid tools have become available, which predict gene functions by using many different data sources together (e.g. MAGIC [Troyanskaya et al. 2003], ProKnow [Pal and Eisenberg 2005], STRING [von Mering et al. 2007]). Unfortunately in most cases the user is not able to see intermediate results and it is not shown whether the underlying databases are up-to-date.

A big bottleneck in the functional annotation is the missing standard for describing the function of a gene or protein. In contrast to sequence or structure information the functional annotation of a protein is written in a human readable fashion. This has many drawbacks like problems with synonyms or the missing relationships between descriptions. Furthermore the human readable description gives no clues how this function was assigned; for example, whether it has been experimentally verified or not or what kind of method was used to predict the function. Ontologies were introduced to solve this problem. However there is not one single comprehensive ontology for all gene classes, but many small and specific ontologies (for example the Enzyme Catalogue (EC) [Webb et al. 1992]) is only adapted for enzymes). To simplify the prediction, most tools deal with only one ontology. But it has been shown that by annotating different kinds of ontology terms to describe the function of a gene, it can be described in a more specific way [Thomas et al. 2007].

[1] *Genes are orthologous, if they were separated by a speciation event. In most cases they share the same function.*

In this chapter a short overview about common ontologies used for function is given. Afterwards different approaches for function prediction are explained. Finally web services and web services workflows are introduced, which are used in this thesis for the integration of different data sources and tools.

a) Function description with ontologies

Often the function of a gene or protein is written in a human readable way. However, because the vocabulary is often invented and reinvented in science, many terms are synonymous [Friedberg 2006]. This makes it hard for humans and machines to interpret it. To make functional annotation accessible to machines, a controlled and well-defined vocabulary is necessary. In the following I will give an introduction to the most frequently used ontologies for function assignment.

Enzyme Commission Classification number (EC): [Webb et al. 1992] This hierarchically organized vocabulary was introduced in 1956 to classify enzymes by the chemical reactions they catalyze. Each enzyme is described by an EC number, which consists of four numbers, separated by dots. While the first three numbers describe the enzyme reaction, the fourth number is used for unique identification. In this process the first number denotes the functional class of the enzyme (Transferase, Hydrolase, etc.), the second and third describe the group of donors or acceptors, which are used by the enzyme.

Gene Ontology term (GO term): [Ashburner et al. 2000] The gene ontology project provides controlled vocabularies to describe genes and gene products in any organism. Three ontologies are publicly available to describe the function of a gene: *Molecular function*, *Biological process* and *Cellular component*. Gene Ontology terms are represented as a directed acyclic graph (DAG) and linked by the two relationships, *is_a* and *part_of*. These relationships enable an easier navigation through the ontology and a faster comparison between the terms. In addition it is possible to add evidence codes[2] to the GO terms, which indicate the method, by which this function has been annotated.

KEGG Ontology term (KO term): [Kanehisa et al. 2004] This hierarchical scheme for orthologous genes was introduced by KEGG to overcome problems with EC numbers and to provide an ontology suited to map genes to regulatory and metabolic pathways. The KEGG ontology was automatically build and manually curated from ortholog clusters of the SSDB database [Sato et al. 2001].

MapMan bin: [Thimm et al. 2004] MapMan bins were developed by the Max-Planck Institute for Plant Physiology to provide a hierarchical system especially suited for plant metabolism. Genes are both mapped automatically and manually by analyzing expression arrays and gas chromatography (GC)/MS metabolite profiles. In addition to that, text search in research papers is used.

FunCat term: [Ruepp et al. 2004] This annotation scheme has a hierarchical, tree-like structure with up to six levels of increasing specificity and is suited for prokaryotes, unicellular eukaryotes, plants and animals. FunCat version 2.1 includes 1362 functional categories of which 28 belong to the main categories, that cover general fields like cellular transport, metabolism and cellular communication/signal transduction.

[2] *http://www.geneontology.org/GO.evidence.shtml*

b) Homology based transfer

Homology-based transfer by database search

This approach uses sequence conservation to transfer the function of a functionally annotated gene to an unknown gene. The most common tool in this field is Blast [Altschul et al. 1997], which searches for homologous sequences in sequence databases. However this method has many drawbacks and makes systemic errors by not accounting for duplication events, evolutionary rate variation, and incorrect annotations. In spite of high sequence conservation the function of the genes can be different [Rost 2002] and gene loss and domain shuffling can lead to wrong annotations, because only part of the putative homologous gene matches to the query sequence very well, which resulted in a high score. These wrong annotations are propagated afterwards through sequence databases [Galperin and Koonin 1998] [Gilks et al. 2002]. Moreover sequence based tools are not in all cases sensitive enough to discover functionally related proteins in other organisms. If the sequence identity drops, as in distantly related organisms, it becomes harder for these tools to detect homologous relationships. In this case and for validation of putative hits it is necessary to check sequences for functionally significant subregions like active sites in enzymes.

Detection of protein domains

Protein domains are conserved parts of a protein structure and sequence, which constitute units of evolution and function. Because most domains are conserved between protein families, they can give clues to the overall function of the protein although no orthologous gene was found by a homology search.

Common methods for protein domain detection are [Durbin et al. 1998]:

Profile Hidden Markov Models (HMMs): Used by HMMER to search HMM databases like PFAM [Bateman et al. 2002]

Profile Specific Scoring Matrices (PSSMs): Included in the Conserved Domain Database (CDD), which can be searched by RPSBlast [Marchler-Bauer et al. 2007]

Regular expressions: Used by PROSITE [de Castro et al. 2006] [Hulo et al. 2006] to search any sequence database

Methods like HMMs are very sensitive, because they allow insertions and deletions. But there is one HMM for each protein domain and each HMM has its own trusted cutoffs. A big bottleneck here is that the seed-alignment of sequences from which the HMM is constructed must be correct. Sequences which do not belong to this domain can decrease the sensitivity and increase the false discovery rate. One of the disadvantages of HMMs is that they are very slow and so this step can be very time-consuming.

In contrast searching with PSSMs is much faster, but not as sensitive as an HMM search. Still, this method is more sensitive than a Blast search.

Regular expressions are used to search domains in sequence databases. Their disadvantages are, that they do not score amino acid frequencies in ambiguous positions and there is no score assigned.

To avoid problems with cutoffs (like in case of HMMs) one can also use so called umbrella databases and tools like InterPro and InterProScan [Mulder and Apweiler 2007], which include many different domain databases and apply tested cutoffs for each method. Each database is searched independently by its own tool. Afterwards all trusted cutoffs are applied and equivalent domains are connected by a unique InterPro identifier. By applying only the trusted cutoffs they achieve a low false discovery rate, but can miss true hits with a lower score. Furthermore some domains are not present in InterPro yet.

The Phylogenomic approach

Phylogenomics considers the evolutionary history of genes to predict functions of uncharacterized genes. A phylogenetic tree is generated from a set of homologous sequences and ontology terms for functional description (see chapter IIIa) are assigned to its leaves (genes). The terms are then transferred within the tree to uncharacterized nodes (uncharacterized genes) by considering speciation and duplication events and branch lengths [Eisen JA et al. 1998].

Because phylogenomics takes the evolutionary history of genes into account, a change of function in a group of paralogous genes can be detected and wrong annotations can be avoided. However, if the number of ontology terms assigned to genes inside the tree is too sparse or the tree is inaccurate, also the phylogenomics approach can lead to wrong annotations.

One of the best performing tools in this area is SIFTER [Engelhardt et al. 2005], which uses a statistical inference algorithm to propagate molecular function Gene Ontology (GO) terms within a phylogenetic tree. Two inputs are required by SIFTER. A reconciled phylogenetic tree and a so-called PLI file, a XML file which includes the gene annotations. SIFTER only uses the lowest level annotated GO terms as candidate functions and is therefore able to assign very specific GO terms to genes. In addition to speciation/duplication events and the branch length between nodes in the tree, SIFTER also considers evidence codes assigned to GO terms. The user of SIFTER can decide which GO terms with which evidence codes should be considered.

For each gene in the tree an initial probability, which is based on the annotated evidence codes (see table 1), is added to each candidate GO term.

GO evidence code	*Description*	*Initial probability*
IEA	Inferred from Electronic Annotation	0.2
IMP	Inferred from Mutant Phenotype	0.8
IGI	Inferred from Genetic Interaction	0.8
IPI	Inferred from Physical Interaction	0.8
ISS	Inferred from Sequence or Structural Similarity	0.4
IDA	Inferred from Direct Assay	0.9
IEP	Inferred from Expression Pattern	0.4
TAS	Traceable Author Statement	0.9
NAS	Non-traceable Author Statement	0.3
RCA	Inferred from Reviewed Computational Analysis	0.4
ND	No biological Data available	0.3
IC	Inferred by Curator	0.4

Table 1: Initial probabilities given by SIFTER for the GO evidence codes.

The initial probability is then changed into a likelihood (see figure 1) and added to the corresponding node in the GO DAG. Afterwards the likelihood for each GO term is combined with a prior value and is down-propagated in the GO DAG to the candidate terms. At the end the likelihood at the candidate GO terms is extracted from the GO DAG and assigned to the gene.

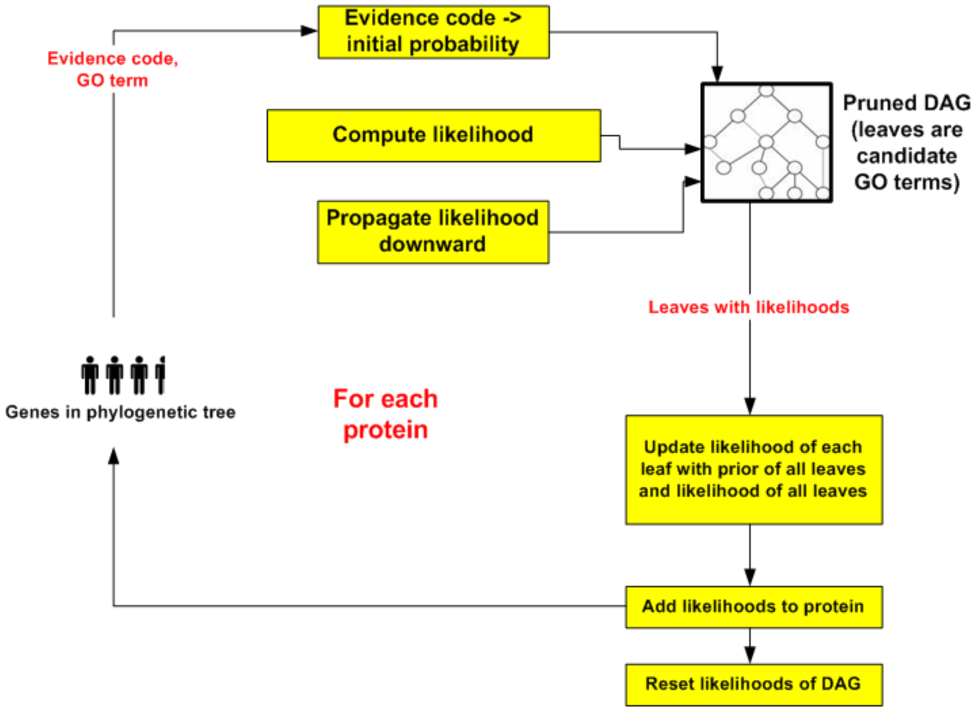

Evidence code -> initial probability
Evidence code, GO term
Compute likelihood
Propagate likelihood downward
Pruned DAG (leaves are candidate GO terms)
Leaves with likelihoods
Genes in phylogenetic tree
For each protein
Update likelihood of each leaf with prior of all leaves and likelihood of all leaves
Add likelihoods to protein
Reset likelihoods of DAG

Figure 1: Setting of the initial likelihood for each candidate GO term to each gene in the phylogenetic tree by SIFTER.

In a second step SIFTER propagates the candidate GO terms to the root of the phylogenetic tree and afterwards back to the leaves of the tree by considering an internally calculated mutation rate. This mutation rate is based on the type of the node (speciation or duplication node) and on the branch length between nodes. The lower the mutation rate is, the more likely it is to transfer functions between nodes. If the node is a duplication node, the mutation rate is increased, because it is assumed that, after a duplication event occurs, the function of one or both genes is modified. The propagation step is based on the approach of Felsenstein et al. [Felsenstein 1981] and is described in [Engelhardt et al. 2006].

c) Chromosomal proximity

Mainly in Prokaryota, but also in other organisms, functionally related genes (e.g. genes working in the same complex [Teichmann and Veitia 2004] or in operons [Blumenthal 2004] [Salgado et al. 2000]) are often placed at the same location on a chromosome. Additionally the gene order is often conserved between related species [Teichmann and Babu 2002]. By comparing these so called "syntenic regions" or shared synteny[3] of related species or individuals, functionally related genes can be identified and functional coherences between genes like physical interactions or activity in the same pathway can be predicted [Poyatos and Hurst 2007] [Enault et al. 2005] [Kolesov et al. 2001].

d) The Rosetta Stone method

The Rosetta stone approach relies on the assumption that in some organisms genes are fused together which in other organism are separate (e.g. α and β subunits of the Trp synthetase in bacteria are fused in fungi [Burns et al. 1990]). By detecting fused genes co-regulation and interaction can

[3] *Shared synteny is defined as the conserved co-localization of genes on chromosomes of related species*

be predicted and functions can be transferred [Date 2008].

e) Phylogenetic profiles

The idea of that approach is that functionally related genes should be co-inherited, because if one gene is lost during evolution the overall function is different due to the lack of the interaction partner. Algorithms in this area cluster profiles of the presence or absence of an orthologous group in a set of species (for review see [Harrington et al. 2008] and [Lee et al. 2007]). However, because the detection of orthologous genes is a crucial step in this analysis and is non-trivial in Eukaryotes, methods in this area are more accurate for prokaryotic genes [Lee et al. 2007].

f) Structural information

In case of low sequence similarity to known genes structural information about the protein can give new insights to the function of the protein of interest, because structure is better conserved than sequence [Brenner et al. 1996] [Rost 1997]. On the one hand that can be done by looking for similar structures of already functionally characterized proteins available in databases like PDB [Kirchmair et al. 2008] and PDBj [Standley et al. 2008] using structure alignment programs (reviewed in [Friedberg et al. 2006]). On the other hand functional critical residues or functionally relevant 3D templates can be identified by looking at the protein stability [Capriotti et al. 2008], the location of the residue in the structure [Kinoshita et al. 2002] [Ng and Henikoff 2006] and the conservation of structural motifs between known proteins [Pazos and Sternberg 2004]. Furthermore structural binding sites can be predicted [Glaser et al. 2006] [Kinoshita et al. 2002] [Ivanisenko et al. 2004] [Golovin and Henrick 2008] [Wei et al. 2007]. For comparison to known binding sites and catalytic sites several public databases are available like the Catalytic Site Atlas (CSA) [Porter et al. 2004] and SCOPEC, a database for catalytic domains [George et al. 2004].

If the 3D structure of the protein is not known the prediction by ab initio programs (for a review and testing of the programs see [Jauch et al. 2007]) can be used as a substitute, or the 2D structure can be used instead.

g) Expression data

Microarray experiments enable the investigation of the expression behavior of many genes together in one experiment. The most common method to predict information about the function of genes from expression data is the clustering of genes based on their expression profile. Information about the function of a known gene in the cluster is then transferred to other genes in the cluster. The hypothesis behind this approach is that genes which are working in the same cellular pathway or interact in some way are required at the same time and are expressed in unison [Boutros and Okey 2005]. Furthermore co-expressed genes can be regulated via one or a few common mechanisms [Boutros and Okey 2005] and by the identification and investigation of gene clusters, hypotheses can be generated about the underlying regulatory mechanism. Of course these hypotheses have to be proven in the lab e.g. by knockdown or knockout experiments.

Many different clustering algorithms exist [Emmert-Streib and Dehmer 2008] [Khatri and Draghici 2005] [Hand and Heard 2005] and all of them have their strengths and weaknesses [Kerr et al. 2008]. To choose an appropriate tool for the clustering process some tools need background knowledge about the underlying data e.g. how many clusters are expected or what is the structure of the cluster [Hand and Heard 2005]. A bottleneck of all algorithms is also the decision on which variables (features) the clustering should be done. If one does not take the right variables, wrong predictions can occur. However, with knowing the data the decision which algorithm to use is biased and therefore the result may not be universally valid. Algorithms like k-means which require the pre-specification of a relatively small number of cluster deal poorly with genes that have no close neighbors in feature space and therefore should ideally form singleton clusters. In this case pattern discovery methods can help, which detect groups of genes which have similar profiles and do not consider the profile shapes of other genes [Hand and Heard 2005].

Besides these problems, there are many other sources of error by investigating expression data like non-complementary binding, varying values because of changed lab conditions, and missing values (e.g. Gene Ontology terms, expression data points). Furthermore, comparing expression data from different Microarrays can also lead to wrong assumptions [Jarvinen et al. 2004] if lab conditions [Irizarry et al. 2005] and the underlying technologies used (e.g. which chip and which normalization is used) are different and the circadian clock had not been considered etc.

Clustering algorithms and pattern discovery methods are mainly useful to predict the biological process of the gene of interest, but should be used with precaution to predict the molecular function of genes. In all cases it is important to look manually at the results afterwards and verify them by running further experiments in the lab.

2) Web services

a) Introduction data integration

Data integration is the cross-association of diverse data organized and presented with a certain purpose. In the current time the amount of biological data increases rapidly and one can find information distributed in several databases. The number of resources that are made available over the web is growing. This creates a need for systems which are able to find the data and to process them in the way that they are not only available at one place, but also combined and collated. There is already a long history of data integration. One of the common examples is the so called data warehouse, in which one extracts data from several data sources and loads them into one database, which then can be queried. But there are some drawbacks about data warehouses, for instance, one needs the space and the compute power to host and integrate all collected data and has to ensure that the data warehouse is always up-to-date.

Another method to face these issues are web services. Web services are software systems that enable the interoperability between two machines in a common network and offer the possibility to compute and/or retrieve data from a distant computer in a machine processable way. Institutes which already provide their data can offer web services to propagate them, so that one does not need to implement a data warehouse. This also means, that it is most important to encourage institutes to make their data publicly available via web services.

b) The BioMOBY project

In 2001 the BioMoby project [The BioMOBY Consortium 2008] was initiated, which also addresses the issues of finding web services and shared data schemata. BioMoby offers a central repository, at which service providers can register their web services and users can find those. Additionally standardized data schemata are offered, which define semantically the input and the output of a service. Normally, a biological web service, which defines a string as input does not give any hint on what kind of string is needed (it can be a protein sequence, a database identifier or a publication abstract). In the BioMoby world a service would be defined for example, as a service which uses a database identifier as input and returns the protein sequence in FASTA format.

3) Web service workflows and reusement of workflows

a) The Taverna project

Taverna [Oinn et al. 2004] is an open source software tool developed by the myGrid team in Manchester for designing and executing workflows. Workflows can be a line-up of web services (see chapter III2) or local tools. Because each component in the workflow is independent, workflows are very flexible.

The common way to execute and develop workflows in Taverna is the use of the standalone application. But workflows can also be executed without using the graphical user interface by interacting with the Taverna application programming interface (API).

b) MyExperiment

To provide an intuitive web interface to find, use and share Taverna workflows the MyExperiment website [Goble and De Roure 2007] was developed by the MyGrid team. Besides uploading, finding, updating and sharing workflows the user of MyExperiment is able to establish new groups and build new scientific communities. Furthermore workflows can be directly loaded, modified and executed in Taverna (reusement of workflows).

IV. Manual annotation and comparison of shared syntenic regions in a hot spot for pathogen resistance in *Solanum tuberosum*, *Solanum demissum* and *Arabidopsis thaliana* to discover new QTLs.

1) Introduction and aims of the project

Solanum tuberosum (cultivated potato) is one of the most important crops of the Solanaceae family. In comparison to the hexaploid wild potato (*Solanum demissum*) the cultivated potato is tetraploid. Both species are non-inbred, annual plants with a genome size between 800 and 1000 megabases and twelve chromosomes. However, cultivated potato genotypes are heterozygous at all ploidy levels, because if the ploidy level is reduced from 4n to 2n the plants become incompatible [Gebhardt et al. 2004]. The whole genome of potato plants is not sequenced yet, but sequencing of both potato and their closest relative tomato (*Solanum lycopersium*) is in progress (see chapter V3).

Because of their agronomic relevance, hot spots for pathogen resistance were identified by comparing RFLP (restriction fragment length polymorphism) linkage maps of the twelve chromosomes [Bonierbale et al. 1988] [Gebhardt et al 1989] [Gebhardt and Valkonen 2001]. One of the identified hot spots is located on potato chromosome V in the region between DNA-based markers GP21 and GP179 [Meksem et al. 2000]. In this region resistance genes and QRLs (quantitative resistance loci) are located for resistance to *Potato Virus X* [De Jong et al. 1997], *Phytophthera infestans* [Leonards-Schippers et al. 1992], *G. rostochiensis* and *G. pallida* [Kreike et al. 1994]. However, only two of them, Rx2 for extreme resistance to *Potato Virus X* [Bendahmane et al. 2000] and R1 for resistance to *Phytophthera infestans* [Ballvora et al. 2002] are functionally characterized. These two genes belong to the superfamily of plant resistance genes, which contain a coiled coil (CC) domain, a nucleotide binding domain and a leucine rich repeat (LRR) domain. The overall sequence identity of genes in this family is not very high. R1 has been introgressed from the wild potato into the cultured potato germ plasm pool.

The aim of this project was to find genes in this hotspot, which can be further examined for function as quantitative trait loci (QTLs), first in silico by functional annotation and then experimentally. In doing this the corresponding region in both haplotypes (~200kbp and ~400kbp) of *Solanum tuberosum* genotype P6/210 was sequenced, manually annotated and genes were functionally characterized. Furthermore syntenic regions in the wild potato *Solanum demissum* [Kuang et al. 2005] and in *Arabidopsis thaliana* were identified and compared.

2) Materials and Methods

a) Sequencing, assembly and gene prediction

BAC clones were sequenced by a company using the shotgun sequencing strategy. The assembly was done first in a company using PreGAP4 and GAP4 from the Staden software package [Krawetz et al. 2003]. After identifying and removing the remaining vector sequences the Megamerger program from the EMBOSS package (Version: 6.0.1) [Rice et al. 2000] was used to merge sequences of overlapping BAC insertions. The assembled contigs were afterwards submitted to the EMBL database [Kulikova et al. 2004] (EMBL accession numbers: R1: EF514212; r1: EF514213)

Different gene prediction tools (GenMark.hmm [Lukashin et al. 1998], FgeneSH [Salamov et al. 2000] and GenomeThreader [Gremme et al. 2005] were run by Remy Bruggmann at the MIPS institute in Munich. Afterwards the APOLLO Genome Annotation Curation Tool (Version: 1.6.4) [Misra et al. 2006] was used to provide a manually annotation for all genes based on the automatic gene prediction of the different tools. Therefore the predicted exons and open reading frames (ORFs) from the gene prediction programs were combined with homologous genes and ESTs found in public databases. The trace files, provided by the company, were manually checked for sequencing errors by using Consed [Gordon et al. 1998], if a putative pseudo gene was discovered.

Further details about materials and methods used for the sequencing and the initial assembly can be retrieved from [Ballvora et al. 2007].

b) Manual functional annotation

Functional annotation was done manually combining homologous genes in the SwissProt database (Release 51) with protein domains and patterns found in the InterPro database (Release 13). For the homolog detection BlastP (Version: 2.2.13) [Altschul et al. 1997] was used. To discover evolutionary relationships inside the disease resistance superfamily a phylogenetic tree from all sequences was build using protpars from the Phylip package [Felsenstein 1993].

c) Shared micro-synteny with the Arabidopsis thaliana and the Solanum demissum genome

Dotter (Version:3.1) and MUMer (Version: 3.18) [Delcher et al. 2002] were used to align and compare the shared syntenic region from different haplotypes of *S. tuberosum* and *S. demissum*. The corresponding BACs in *S. demissum* had been taken from [Kuang et al. 2005]. For the identification of syntenic genes in *A. thaliana* Inparanoid (Version: 1.35) [O'Brien et al. 2005] in combination with BlastX (Version: 2.2.13) [Altschul et al. 1997] against the TAIR6 database [Weems et al. 2004] were used. We defined shared syntenic blocks by the criterion, that at least three orthologous genes are identified within a roughly comparable physical distance on the chromosome. Because of their limited information value transposons and resistance genes were excluded from the comparison.

3) Results

The assembly of 743,152 kbp of genomic sequence from seven R1 and three r1 BAC insertions results two distinguished and unambiguous DNA contigs with a length of 417,445 kbp and 202,781 kbp of R1 and r1.

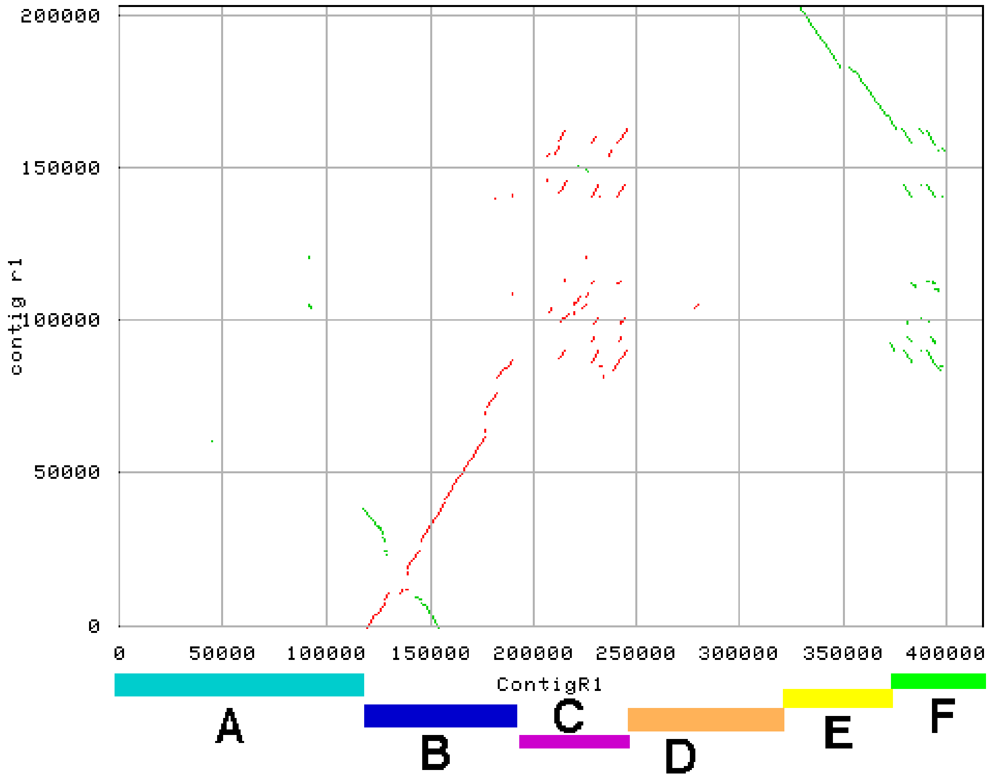


Figure 2: Comparison of the haplotypes R1 with r1 of Solanum tuberosum. The figure was taken from Ballvora et al. 2007.

R1 and r1 share a high conserved region (see region **B** and **E** in figure 2), which is interrupted and extended by two hyper variable regions, which were non-alignable (see region **C,D** and **F** in figure2). All identified resistance genes in R1 (8 genes) and r1 (4 genes) are located in this hyper variable region. One of them in R1 (gene between 20 and 21 (see figure 3)) and one gene in r1 (gene 49 (see figure 3)) seem to be incomplete (putative pseudo genes), because they were disrupted by transposons. The first gene of R1 (gene 1 in figure 3) could also not be well annotated, because there have not been more sequences available at this side of the R1 contig and the corresponding region (region **A** in figure 2) was not obtained from r1. Region **B** (see figure 2) shows a palindromic structure, which corresponds to an inverted repeat of two RNA-directed RNA polymerases, which were separated by hypothetical protein and a retro-transposon. In R1 four and in r1 one of these hypothetical proteins could be identified, but no expression data was available to confirm their transcription. However all hypothetical proteins seem to belong to the same protein family. Blast and profile searches only returned hits to potato proteins, which support the idea that it is an unknown potato specific protein family. We could also identify a genomic inversion in region **E** (see figure 2), which is highly conserved (sequence identity: 99%) and contains 5 genes (gene 43–38 in figure 3).

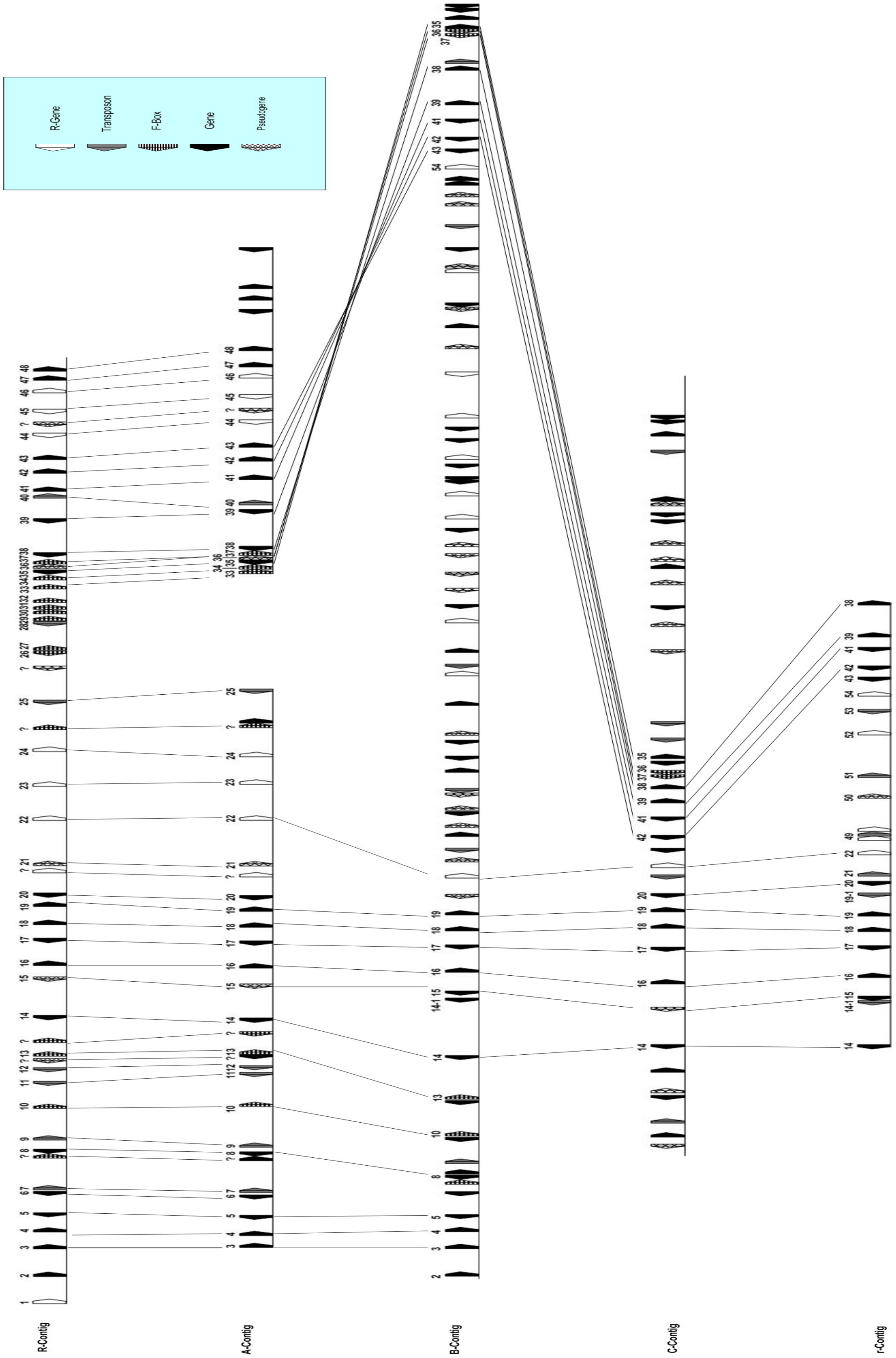


Figure 3: Comparison between R1 and r1 from S. tuberosum and haplotypes A, B and C from S. demissum. The figure was taken from Ballvora at al. 2007.

We could manually annotate 55 genes on R1 and 22 genes on r1. 6 genes on R1 and r1 are putative transposon genes and 5 genes on R1 and one gene on r1 are putative pseudo genes. 14 genes on R1 were annotated as F-Box genes, but none of these were found in r1. A complete list of all annotated genes is available in table 2. Except for the transposons most genes are conserved between R1 and r1 in sequence, order and orientation. Furthermore I had some difficulties to identify syntenic genes between R1 and r1 of the disease resistance family, because these genes seems to mutate very fast. A sequence comparison and an analysis of a phylogenetic tree with all members found in *S. tuberosum* showed no result. I could identify R1 on contig R1 as gene 44 (see figure 3). Because the neighboring genes 43 and 42 are included in the inversion and gene 44 and his neighbor at the other side are conserved in order and orientation with gene 54 and 52 on r1, I assume, that the inversion also includes these two genes. The inversion ends after gene 45, because the next genes are not similar in order and orientation.

a) Shared micro-synteny with Solanum demissum

Comparison of the corresponding haplotypes A, B and C in *S. demissum* revealed similar features [Kuang et al. 2005]. We found, that the A contig in *S. demissum* shows a high sequence identity (99%) with the R1 contig in *S. tuberosum*. Contig B and C are more similar to contig r1 in *S. tuberosum* except for the hyper variable region, which is much larger on the B contig and not included on the C contig. Because [Kuang et al. 2005] did not sequence the region between gene 25 and 35 on contig A, they were not able to detect the inversion, which is also present between contig A and contig B and C.

b) Shared micro-synteny with Arabidopsis thaliana

We identified five syntenic blocks in *A. thaliana* (see table 1). The largest syntenic region found spans almost the complete R1 contig and 54 kbp on chromosome 1 in *A. thaliana* and includes seven genes of *S. tuberosum.* Five of these genes are conserved in sequence, order and orientation and two of them (gene 38 and 43) show reverse order and orientation compared with *A. thalina*. These genes are included in the inversion between R1 and r1, so gene 38 and 43 on r1 have the same order and orientation as in *A. thaliana*. This is the same case for AT1G14270.1 until AT1G14300.1, where AT1G14270.1 (gene 17 on R1) and AT1G14280.1 (gene 19 on R1) are in the same order and orientation as on R1 and AT1G14290.1 and AT1G14300.1 are in reverse order on contig R1, but included in the inversion. Non of the syntenic genes are disease resistant genes and all disease resistant genes in potato have the highest sequence similarity to *RPP13*. RPP13 confers resistance to *Peronospora parasitica* and is located on chromosome 3 outside any detected syntenic region.

Syntenic block	***S. tuberosum ORF***	***A. thaliana ORF***	***A. thaliana BAC***	***A. thaliana ORF position [Mbp]***	***A. thaliana block size***	***S. tuberosum block size***
I	ORF17	AT1G14270.1	F14L17	4875	7 kbp	215 kbp
	ORF19	AT1G14280.1	F14L17	4878		
	ORF43	AT1G14290.1	F14L17	4880		
	ORF41	AT1G14300.1	F14L17	4882		
II	ORF4	AT1G26880.1	T2P11	9316	18 kbp	345 kbp
	ORF5	AT1G26870.1	T2P11	9313		

Syntenic block	***S. tuberosum ORF***	***A. thaliana ORF***	***A. thaliana BAC***	***A. thaliana ORF position [Mbp]***	***A. thaliana block size***	***S. tuberosum block size***
	ORF18	AT1G26850.1	T2P11	9301		
	ORF42	AT1G26840.1	T2P11	9298		
III	ORF2	AT1G69600.1	F24J1	26168	54 kbp	405 kbp
	ORF3	AT1G69610.1	T6C23	26190		
	ORF4	AT1G69620.1	T6C23	26193		
	ORF43	AT1G69640.1	T6C23	26197		
	ORF38	AT1G69690.1	T6C23	26221		
	ORF47	AT1G69700.1	T6C23	26224		
	ORF48	AT1G69710.1	T6C23	26226		
IV	ORF2	AT3G28920.1	MYI13	10941	106 kbp	25 kbp
	ORF4	AT3G28900.1	K5K13	10904		
	ORF5	AT3G29035.1	K5K13	11035		
V	ORF2	AT5G39760.1	MKM21	15928	28 kbp	25 kbp
	ORF3	AT5G39785.1	MKM21	15946		
	ORF5	AT5G39820.1	MKM21	15956		

Table 1: Syntenic regions between S. tuberosum and A. thaliana. Table was taken from Ballvora et al. 2007.

4) Discussion

In this project a complete sequencing pipeline was run manually to detect problems with fast evolving genes and to improve the structural and functional annotation. In genome projects this approach is not feasible, because it is too time-consuming and so it has to be done automatically. However it can be further improved if parts like the gene annotation are manually verified.

Syntenic regions in both *S. demissum* and *A. thaliana* could be identified, which give new insights to the evolutionary history of *S. tuberosum* strain P6/210. We assume, that the R1 contig is introgressed from *S. demissum* into *S. tuberosum*, because the overall sequence identity is very high and all genes are conserved in sequence, order and orientation. The r1 contig seems to originate from either *S. tuberosum* or *S. spegazzinii*, based on the notation that the parental donor of the r1 homolog was an interspecific hybrid between *S. tuberosum* or *S. spegazzinii* [Barone et al. 1990]. In *A. thaliana* five syntenic blocks were identified, which include genes in the same order and orientation to R1 and in reverse order to R1. Genes in reverse order to R1 are in an inverted region of ca. 70 kbp, which could be identified by comparing R1 and r1 and are therefore in the same order and orientation compared to r1. We assume, that the inversion occurred in the R1 linkage after the divergence of *Arabidopsis* and the *Solanum* species. By comparing *S. tuberosum* and *S. demissum* also well conserved regions could be identified, which are separated and extended by a hyper variable region. No conservation in this hyper variable region between all potato haplotypes could be detected and no syntenic region could be identified in *A. thaliana*. Genes in this region are mostly disease resistance genes, transposons and F-Box genes, which show a fast mutation rate. This fast mutation and duplication rate is important for potato plants to adapt to pathogens. Because of the fast mutation rate, I was not able to identify subgroups in the disease resistance family or to find any hints to which pathogens the resistance gene is directed.

As most likely candidates for QTLs I identified 14 F-Box genes. F-box proteins are involved in various signaling pathways in *Arabidopsis thaliana* and a F-box domain was identified in the SGT1 protein that was shown to play a role as co-chaperone in the stabilization of R-proteins [Shirasu and Schulze-Lefert 2000] [Schulze-Lefert 2004]. To confirm this QTL, further investigations are needed.

Also the new potato specific protein family is interesting. Here again the functional annotation could not give any clues to the function of these genes, because they are not similar to any functionally characterized gene. This family could be a new kind of transposon, which is only present in potato plants, but no transposon specific domains or motifs could be identified. Unfortunately there was no expression data available, which could be used to confirm the transcription of these genes. The expression data would also be useful to confirm putative pseudo genes.

V. Automatic annotation in genome projects

This chapter shows how to facilitate the functional characterization of genes for whole genomes. In this case an automatic pipeline is needed, because manual functional annotation or running experiments for all predicted genes is not feasible. I introduce a phylogenomic pipeline for the automatic functional annotation of molecular function Gene Ontology terms, which uses SIFTER , one of the best performing phylogenomic tools [Engelhardt et al. 2005]. This is tested in the ongoing *Medicago truncatula* genome project. Based on these results the pipeline is improved and applied on the genome of *Sorghum bicolor* and on the first available part of the tomato genome. Furthermore specific gene families, which are unknown in plants so far, are extracted and functionally characterized. Especially the Transferrin protein family is investigated further, because it is a putatively old gene family which is well known in animals, insects and algae, but not yet known in higher plants.

1) Introduction

The *Medicago truncatula* genome project

Medicago truncatula or barrel medic (see image at the left side) has a small diploid genome with eight chromosomes. It is self-fertile and has a rapid generation time and prolific seed production[4]. Other advantages are that it is amenable to genetic transformation and large collections of mutants and ecotypes are available[5]. Because of these attributes, *M. truncatula* has been chosen as the new model organism for legumes.

Most legumes (or Fabaceae) live in a symbiotic relationship with bacteria. These bacteria (or rhizobia) live in their roots within structures called root nodules and have the ability to take nitrogen gas out of the air and convert it to a form of nitrogen that is usable to the host plant. This process is called “nitrogen fixation” and reduces fertilizer costs for farmers and gardeners, because they use legumes in a crop rotation to replenish soil that has been depleted of nitrogen[6]. *M. truncatula* lives in a symbiotic relationship with the rhizobia *Sinorhizobium meliloti* and arbuscular mycorrhizal fungi. This makes *M. truncatula* an interesting object to study symbiotic relationships between plants, bacteria and fungi.

The sequencing of the *M. truncatula* genome started in 2003. Six chromosomes are sequenced in the USA and two chromosomes are sequenced in Europe. To provide a high quality automated gene prediction and annotation for all finished sequences generated by the Medicago genome sequencing project the International Medicago Genome Annotation Group (**IMGAG**) was initiated.

[4] *http://mips.gsf.de/proj/plant/jsf/medi/index.jsp*

[5] *http://en.wikipedia.org/wiki/Medicago_truncatula*

[6] *http://en.wikipedia.org/wiki/Fabaceae*

In October 2006 about 60% of the gene-rich euchromatin was sequenced and a first sequence release was made (Mt1.0). The gene calling process was complete and 43616 genes were annotated. All genes had a human readable description assigned by using the most significant hit in the InterPro database using InterProScan and, if there was no significant hit in InterPro, the most significant Blast hit against the TIGR database [Lee and Quackenbush 2003] [Chan et al. 2007] was used.

Goal of this project was the assignment of molecular function Gene Ontology (GO) terms to as many protein coding genes of *M. truncatula* as possible, at the same time avoiding wrong annotations. In this process the assigned GO terms should be as specific as possible. For this approach SIFTER (see chapter III1b) was tested, which uses a statistical inference algorithm to propagate molecular function GO terms within a phylogenetic tree. Engelhardt et al. claimed in their paper that SIFTER achieved an accuracy of 96% using their test set and should be able to assign very specific terms.

To apply SIFTER to all annotated Medicago genes a new pipeline had to be implemented. This pipeline provides a phylogenetic tree and a so called "*PLI-file*", which includes all homologous genes in the tree and the assigned GO terms with the corresponding evidence codes available to each gene. No pipeline is provided in the SIFTER package, but several Perl scripts to build a tree from a PFAM alignment and to build the PLI file using the SwissProt database. However, if the genes in the tree are not included in the SwissProt database or the query protein has no domain present in the PFAM database, these scripts can not be used.

To validate GO terms and increase the number of functionally annotated genes at the end of the analysis InterProScan in combination with InterPro2GO from the GOA project [European Bioinformatics Institute 2008] was also used to assign GO terms to all predicted Medicago genes. To run two programs in parallel has the advantage that the results can be compared at the end to find out which program performs better in the number of annotated genes and whether it useful to run both programs instead of one.

Another goal of the project was to find non-plant specific genes in the Medicago genome, which can be used as candidate genes for further experiments. Genes, which have homologous genes in animals, but not in plants were identified. One interesting protein family, the Transferrin family, was found and investigated further.

The *Sorghum bicolor* genome project

The diploid, annual, in some cultivars perennial plant *Sorghum bicolor* (see image at the left side) belongs to the monocotyledonous green plants in the family Gramineae (Poaceae) (Common name: Grasses). Sorghum has 10 chromosomes and a total genome size of approximately 770Mb.

Sorghum was sequenced by the Sorghum Genome Project at the DoE Joint Genome Institute and will be published in 2009. The gene calling process is complete and 27458 genes were predicted and are publicly available[7] under the Fort Lauderdale genome data release policy. These sequence data were produced by the US Department of Energy Joint Genome Institute http://www.jgi.doe.gov/ in collaboration with the user community. For none of these genes, functional annotations are available (status: Beginning of 2008).

[7] *http://www.phytozome.net/Sorghum*

Thus the *Sorghum bicolor* genome project is suitable to test and improve the automatic functional annotation pipeline to provide an accurate functional annotation for the Sorghum genome. Goal was on the one hand to provide a flexible and accurate automatic functional annotation pipeline, which can be used by other genome projects. In this context the pipeline used in the Medicago genome project has been further improved and tested on the Sorghum genome. On the other hand functional annotation results should be compared with functional annotations of other genomes to find conspicuities in the Sorghum genome.

The *Solanum lycopersicum* (tomato) genome project

Figure 4: Harvesting of tomato fruits for research. This figure is taken from http://www.eu-sol.net/

Solanum lycopersicum, better known as tomato, is a perennial, herbaceous plant and belongs to the Solanaceae family (nightshade). Tomato as well as other members of the Solanaceae family have evolved from South America and were brought 1498 by Christopher Kolumbus to Spain and Portugal[8]. Until then the plants have been cultivated mainly in Spain, Portugal, the Netherlands and Italy and tomato is now one of the most popular vegetables in Europe. Members of the Solanaceae family like potato, physalis, tobacco and so on are closely related to each other and show a high sequence conservation.

Because of its small genome size of about 950 Mb, tomato is used as a new model organism for the Solanaceae family. Furthermore there is strong interest in improving fruit quality by extracting genes, which are involved in disease resistance, plant architecture and the nutritional value, taste, flavor, fragrance and starch composition of the fruit[9]. The sequencing of the euchromatic part of the genome (~250 Mb), which is done by eleven countries around the world, is underway. The annotation effort (gene finding and functional annotation) is done by the International Tomato Annotation Group (ITAG) using the pipeline in figure 5. All ab initio gene finders are trained with tomato data to provide the most accurate prediction. My part in the pipeline was the prediction of Gene Ontology terms by using the phylogenomic pipeline with SIFTER in combination with InterProScan and InterPro2GO.

To provide the annotation of the genes as soon as their corresponding sequence is sequenced by the sequencing centers and publicly available in databases like GenBank [Sayers et al. 2008] and SGN [Mueller et al. 2005], the ITAG annotation pipeline is run iteratively in batches of available tomato sequence data. The first batch (batch11) for which the complete pipeline was run was available in May 2008 and consists of 283 contigs. 9942 genes were predicted.

[8] *http://www.economy-point.org/t/tomato.html*
[9] *http://www.eu-sol.net/science*

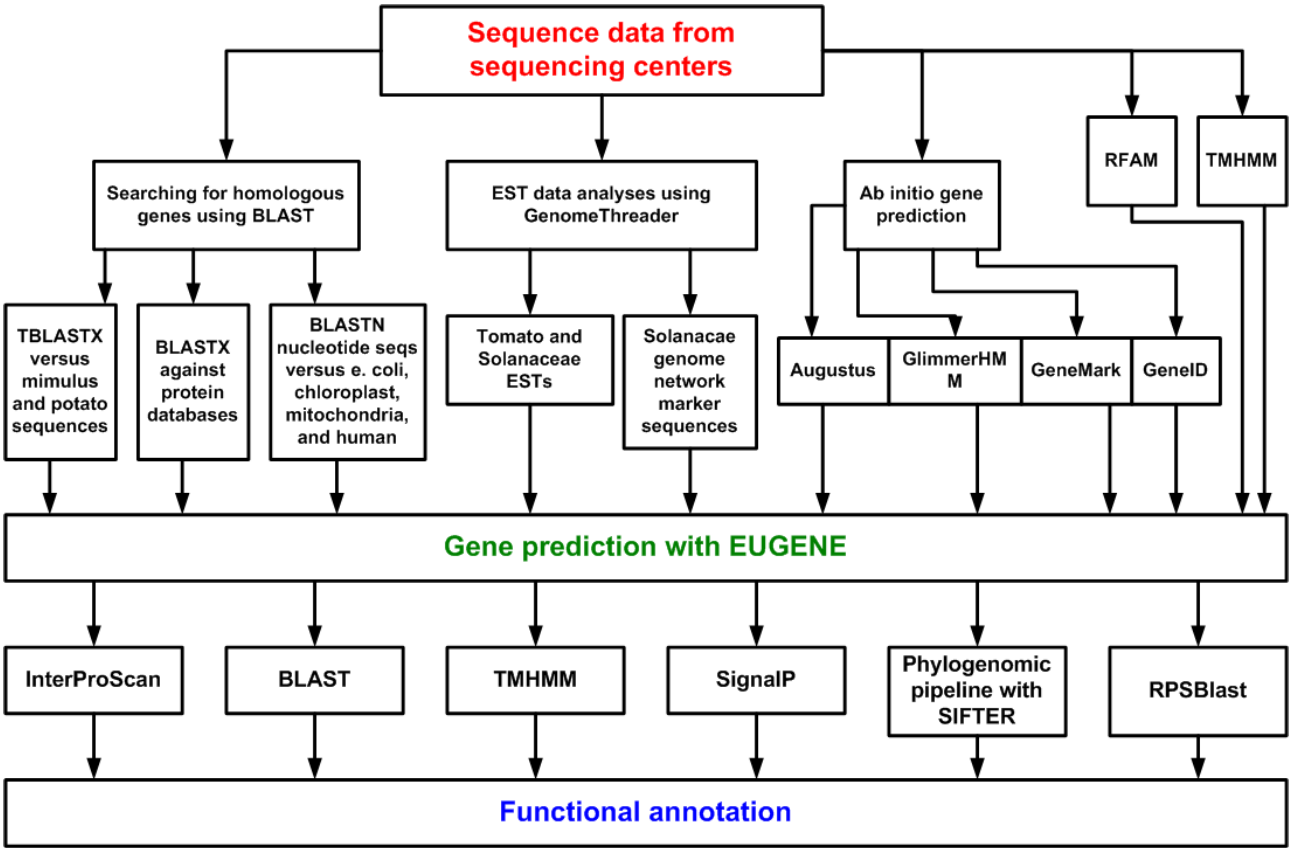

Sequence data from sequencing centers
Searching for homologous genes using BLAST
EST data analyses using GenomeThreader
Ab initio gene prediction
RFAM
TMHMM
TBLASTX versus mimulus and potato sequences
BLASTX against protein databases
BLASTN nucleotide seqs versus e. coli, chloroplast, mitochondria, and human
Tomato and Solanaceae ESTs
Solanacae genome network marker sequences
Augustus
GlimmerHMM
GeneMark
GeneID
Gene prediction with EUGENE
InterProScan
BLAST
TMHMM
SignalP
Phylogenomic pipeline with SIFTER
RPSBlast
Functional annotation

Figure 5: ITAG pipeline. Sequence data coming from the different sequencing centers is used as input for different gene finders, Blast searches [Altschul et al. 1997] against diverse protein and nucleotide databases, GenomeThreader analysis [Gremme et al. 2005] against EST collections, RFAM [Gardner et al. 2009]and TMHMM [Krogh et al. 2001]. All this information goes into EUGENE [Schiex et al. 2001] to predict the gene exon/intron structure. The translated amino acid sequence is then used as input in several functional prediction tools to perform an accurate functional annotation.

2) Material & Methods

a) Homolog detection

Homolog detection used in the Medicago pipeline

To enable a fast search for homologous proteins, which can be used as candidates to build a phylogenetic tree, Blastp was used. To further speed up the analysis Blastp was run against a database of proteins, which have an experimentally verified or reviewed GO term assigned. Because there was no such database available, gene identifiers from 7 Eukaryota gene association files available at the Gene Ontology website [Gene Ontology 2008] and from the gene association file provided by the GOA project [Camon et al. 2004] were extracted, selecting only those which have an experimentally verified or reviewed GO term assignment (not IEA and ND) (see table 2).

These identifiers were mapped to the corresponding amino acid sequences via FASTA files or web services (see table 2) available from the institutes, which uploaded the corresponding gene association files to the Gene Ontology website. At the end all extracted amino acid sequences were merged and provided as a Blast database. Additionally all sequences with their corresponding identifiers and GO term annotations (including evidence codes) were stored in the AFAWE MySQL database (see figure 22) to increase the performance of the pipeline once more.

$$Overlap = min\left(\frac{(query_{end} - query_{start})}{query_{length}}, \frac{(hit_{end} - hit_{start})}{hit_{length}}\right)$$

Equation 1: Overlap computation between Blast query sequence and Blast hit sequence

To extract the candidate homologous genes from the Blast result an overlap cutoff of at least 70% and an e-value cutoff of smaller than 1 was applied to all Blast hits. The overlap was computed by equation 1.

Organism	***Annotation file***	***Date***	***Identifier used***	***Name of fasta file or web service used (Database)***
Arabidopsis thaliana	gene_association.tair	15/09/2006	AGI locus code	TAIR6 (TAIR)
Saccharomyces cerevisiae	gene_association.sgd	15/09/2006	SGI ID	orf_trans.fasta (SGD)
Drosophila melanogaster	gene_association.fb	19/08/2006	FlyBase ID	dmel-all-gene-r4.3.fasta (FlyBase)
Caenorhabditis elegans	gene_association.wb	26/08/2006	WormBase ID	current.tar.gz (WormBase)
Oryza sativa	gene_association.gramene	27/08/2006	UniProt identifier	DBFetch (UniProt)
Candida albicans	gene_association.cgd	15/09/2006	CGD ID	orf_trans_all_assembly_20.fasta
Dictyostelium discoideum	gene_association.dictyBase	15/09/2006	dictyBaseID	dicty_curated_models_protein (DictyBase)
Different organisms	gene_association.goa_uniprot	15/09/2006	UniProt identifier	DBFetch (UniProt)

Table 2: Gene association files downloaded from Gene Ontology and files/web services used to build a database of proteins, which have an experimentally verified or reviewed GO term assignment.

Improved homolog detection

To improve the prediction accuracy of SIFTER a complete and accurate phylogenetic tree is needed. This needs to include a comprehensive gene neighborhood for the query gene. However, because the pipeline should be suitable for large genome sets, it needs to be reasonably fast.

I improved the former homolog detection in two ways. Firstly the Blast database used in the former pipeline was replaced by a database, which only includes completely sequenced organisms from organisms, for which also GO term annotations were available. The complete genome of all these organisms (see table 3) was downloaded from the RefSeq download page[10] (release 21 from 15-01-2007), integrated in the AFAWE MySQL database (see figure 22) and provided as a protein Blast database by using formatdb [Altschul et al. 1997]. In addition all gene association files provided by the Gene Ontology consortium were downloaded (Download date: 14-08-2007) and included in the AFAWE MySQL database. To map genes from RefSeq to the corresponding GO terms provided in the gene association files, mapping files from different resources (see table 3) were downloaded and included in the AFAWE MySQL database. For organisms for which no mapping between organism database identifier and RefSeq protein identifier was available a Blast search was used to find 100% identical and 100% overlapping sequences between the RefSeq and the model organism databases. Also these mappings were integrated in the AFAWE database to map as many genes as possible to

[10] *ftp://ftp.ncbi.nih.gov/refseq/release/*

GO terms.

Organism	***Gene association file***	***Mapping file or fasta file used (Database)***	***Identifier mapping used***	***Download date***
A. phagocytophilum	gene_association.tigr_Aphagocytophilum	protein_tigr_annotation_20070814.fasta (TIGR)	TIGR_CMR ID → Blast against RefSeq	14-08-2007
A. thaliana	gene_association.tair	TAIR7_NCBI_mapping_prot (TAIR)	AGI locus code → RefSeq ID	25-04-2007
B. anthracis	gene_association.tigr_Banthracis	protein_tigr_annotation_20070814.fasta (TIGR)	TIGR_CMR ID → Blast against RefSeq	14-08-2007
B. taurus	gene_association.goa_cow	Iproclass.tb (PIR) & ipi.BOVIN.xrefs.gz (IPI)	UniProt ID → RefSeq ID	14-08-2007
C. elegans	gene_association.wb	Wormpep179 (WormBase)	WormBase ID → Blast against RefSeq	14-08-2007
C. jejuni	gene_association.tigr_Cjejuni	protein_tigr_annotation_20070814.fasta (TIGR)	TIGR_CMR ID → Blast against RefSeq	14-08-2007
C. albicans	gene_association.cgd	orf_trans_all_assembly_20.fasta (CGD)	CGD ID → Blast against RefSeq	01-12-2006
C. burnetii	gene_association.tigr_Cburnetti	protein_tigr_annotation_20070814.fasta (TIGR)	TIGR_CMR ID → Blast against RefSeq	14-08-2007
D. rerio	gene_association.zfin	mapping_ZFIN_RefSeq.txt (ZFIN)	ZFIN ID → RefSeq ID	14-08-2007
D. ethenogenes	gene_association.tigr_Dethenogenes	protein_tigr_annotation_20070814.fasta (TIGR)	TIGR_CMR ID → Blast against RefSeq	14-08-2007
D. melanogaster	gene_association.fb	Dmel-all-gene-r5.2.fasta (FlyBase)	FlyBase ID → Blast against RefSeq	17-08-2007
E. chaffeensis	gene_association.tigr_Echaffeensis	protein_tigr_annotation_20070814.fasta (TIGR)	TIGR_CMR ID → Blast against RefSeq	14-08-2007
G. gallus	gene_association.goa_chicken	Iproclass.tb & ipi.CHICK.xrefs.gz	UniProt ID → RefSeq ID	14-08-2007
G. sulfurreducens	gene_association.tigr_Gsulfurreducens	protein_tigr_annotation_20070814.fasta (TIGR)	TIGR_CMR ID → Blast against RefSeq	14-08-2007
H. sapiens	gene_association.goa_human	Iproclass.tb & ipi.HUMAN.xrefs.gz	UniProt ID → RefSeq ID	14-08-2007
L. monocytogenes	gene_association.tigr_Lmonocytogenes	protein_tigr_annotation_20070814.fasta (TIGR)	TIGR_CMR ID → Blast against RefSeq	14-08-2007
M. capsulatus	gene_association.tigr_Mcapsulatus	protein_tigr_annotation_20070814.fasta (TIGR)	TIGR_CMR ID → Blast against RefSeq	14-08-2007
M. musculus	gene_association.mgi	MRK_SwissProt_TrEMBL.rpt (MGI) & Iproclass.tb (PIR)	MGI ID → UniProt ID → RefSeq ID	17-08-2007
N. sennetsu	gene_association.tigr_Nsennetsu	protein_tigr_annotation_20070814.fasta (TIGR)	TIGR_CMR ID → Blast against RefSeq	14-08-2007
O. sativa	gene_association.gramene_oryza	Iproclass.tb (PIR)	UniProt ID → RefSeq ID	14-08-2007

Organism	***Gene association file***	***Mapping file or fasta file used (Database)***	***Identifier mapping used***	***Download date***
P. aeruginosa	gene_association.pseudocap	pseudomonas_aeruginosa_PAO1_2007-06-19.fasta (PseudoCAP)	PseudoCAP ID → Blast against RefSeq	19-06-2007
P. syringae	gene_association.tigr_Psyringae	protein_tigr_annotation_20070814.fasta (TIGR)	TIGR_CMR ID → Blast against RefSeq	14-08-2007
R. norvegicus	gene_association.rgd	GENES_RAT (RGD)	RGD ID → RefSeq ID	14-08-2007
S. cerevisiae	gene_association.sgd	orf_trans.fasta (SGD)	SGD ID → Blast against RefSeq	14-08-2007
S. pompe	gene_association.GeneDB_Spompe	pompep (Sanger GeneDB)	GeneDB_Spombe ID → Blast against RefSeq	14-08-2007
S. oneidensis	gene_association.tigr_Soneidensis	protein_tigr_annotation_20070814.fasta (TIGR)	TIGR_CMR ID → Blast against RefSeq	14-08-2007
S. pomeroyi	gene_association.tigr_Spomeroyi	protein_tigr_annotation_20070814.fasta (TIGR)	TIGR_CMR ID → Blast against RefSeq	14-08-2007
V. cholerae	gene_association.tigr_Vcholerae	protein_tigr_annotation_20070814.fasta (TIGR)	TIGR_CMR ID → Blast against RefSeq	14-08-2007

Table 3: Mapping between GO terms and RefSeq identifiers via database identifiers from the gene association files from Gene Ontology. The download date is the date when the mapping file respectively the protein fasta file have been downloaded.

Secondly an iterative Blast search (parameter: -e 1 -F F -m 8) against the RefSeq database is used and only putative orthologs and in-paralogs are extracted from the result instead of using a definite cutoff. Putative orthologs are defined as the first hit of each organism in the Blast result, if it has an overlap greater than 60% between query and hit. The score of this alignment is called "ortholog score". Afterwards a Blast search is run with each candidate orthologous gene as query and all hits with a score greater or equal than the ortholog score are defined as further candidate orthologs and in-paralogs. The sequences and identifiers of these proteins are then stored together with the original candidate orthologs from the first Blast search and the original query protein in a fasta file.

b) Pipeline implementations to assign Gene Ontology terms to genes

Pipeline used in the Medicago genome project

The pipeline used in the Medicago genome project to annotate GO terms to genes consists of two parts. One part includes a workflow which first searches for putative homologous genes (see chapter V2a), aligns their amino acid sequences using MUSCLE [Edgar et al. 2004], builds a phylogenetic tree using QUICKTREE [Howe et al. 2002], reconciles the phylogenetic tree with a species tree from the NCBI taxonomy database [Sayers et al. 2008] using FORESTER [Zmasek and Eddy 2001] and runs SIFTER (version 0.3) to propagate molecular function GO terms within the tree. To provide a PLI file as input for SIFTER, which includes genes and their corresponding GO terms, GO terms for all putative homologous genes were extracted from the AFAWE MySQL database (see figure 22). At the end of the workflow the predicted molecular function GO term with the highest score for the query gene are extracted from the SIFTER output and assigned to the query gene (this is the default setting in SIFTER).

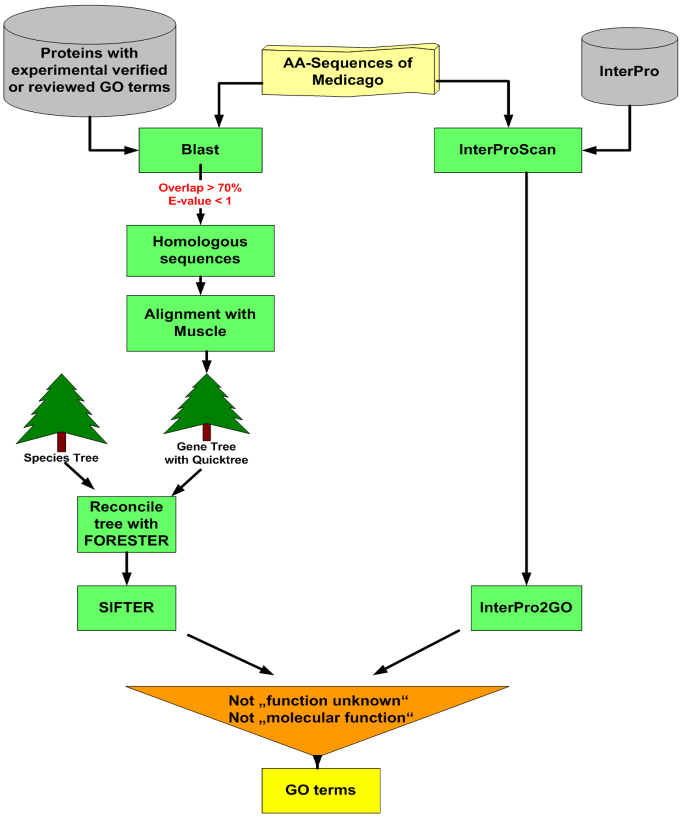


Figure 6: Pipeline for the automatic annotation of M. truncatula genes.

The other part of the pipeline runs InterProScan to look for protein domains in the query sequence. Afterwards InterPro2GO is applied to map GO terms. For this approach InterProScan results for all Medicago proteins computed at the JCVI institute (formally TIGR) were stored in the AFAWE MySQL database. In addition the InterPro2GO file was downloaded from the Gene Ontology website[11] and GO terms for each found domain were stored in the same database.

Results from both SIFTER and InterProScan were filtered by excluding GO terms for “molecular function” and “function unknown”, because the information content of these terms is too low. Afterwards the results from the SIFTER pipeline and the InterProScan approach were compared to find out whether there is any overlap between them and which tool performs better in specificity and number of annotated proteins. At the end the results were joined by assigning only the most specific GO terms returned from any of the two tools.

Improvement of the automatic functional annotation pipeline and implementation of a web service workflow

The SIFTER pipeline used in the Medicago genome project was improved in several ways to build a more comprehensive phylogenetic tree, which is used as input for SIFTER (see figure 7). On the one hand the homolog detection was improved (see chapter V2a) to get a most comprehensive gene neighborhood, on the other hand tools to build a multiple alignment between all homologs and to create a phylogenetic tree were replaced by more accurate tools.

Instead of MUSCLE, MAFFT [Katoh et al. 2005] is used in the new pipeline to build a more accurate and fast multiple sequence alignment from all discovered inparalogs and orthologs [Nuin et al. 2006] [Perrodou et al 2008]. By filtering out columns in the alignment with more than 60% gaps only conserved regions are extracted from the alignment. To speed up the phylogenetic tree building for big trees, but be as accurate as possible two different approaches are used to build the phylogenetic tree. If there are less than 20 proteins in the alignment PHYML [Guindon and Gascuel 2003], a fast maximum likelihood approach is used. For more than 20 sequences BIONJ [Gascuel 1997], a fast and accurate neighbor joining approach, is run. To reconcile the phylogenetic tree with a species tree and assign duplication and speciation nodes FORESTER [Zmasek, 2001] is used. SIFTER in version 0.3 is replaced by SIFTER version 1.2, which is much faster than the first version. Furthermore the SIFTER source code (version 1.2) was modified so that always the best three GO terms are returned for each gene in the tree instead of only one GO term.

[11] *http://www.geneontology.org/external2go/interpro2go*

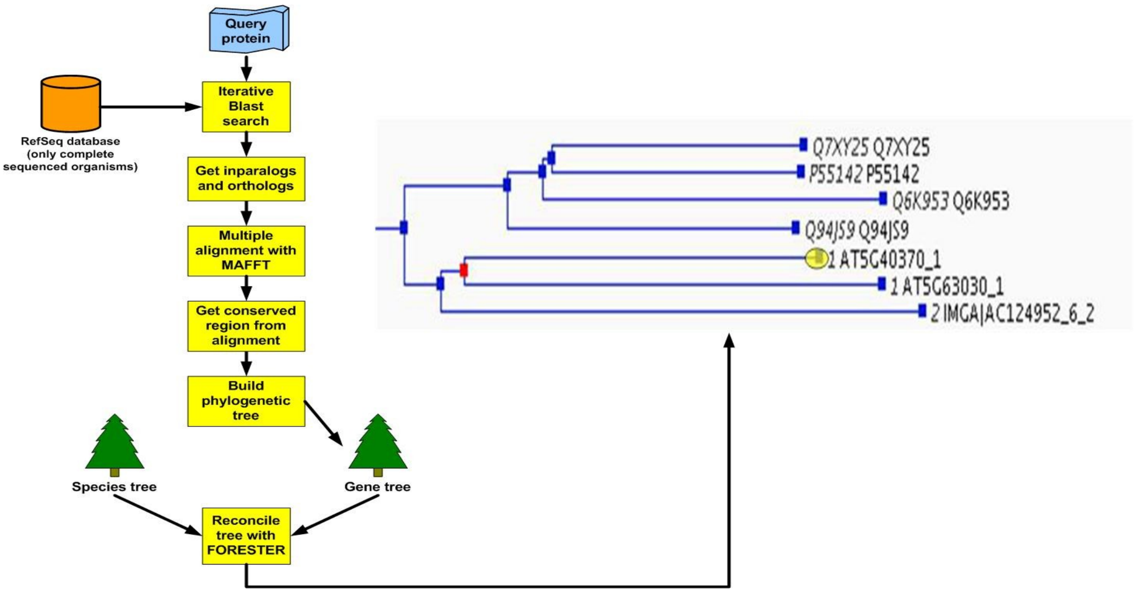


Figure 7: New SIFTER pipeline. The goal was to improve the phylogenetic tree which is used as input for SIFTER.

To make the pipeline usable for other genome projects and make parts of it easily exchangeable, each step in the pipeline is implemented as a BioMOBY web service [The BioMOBY Consortium 2008], except for FORESTER and SIFTER, which are combined in a single web service. Furthermore a Taverna workflow (see figure 8) was built and is publicly available at the MyExperiment website (*http://www.myexperiment.org/tags/638)*.

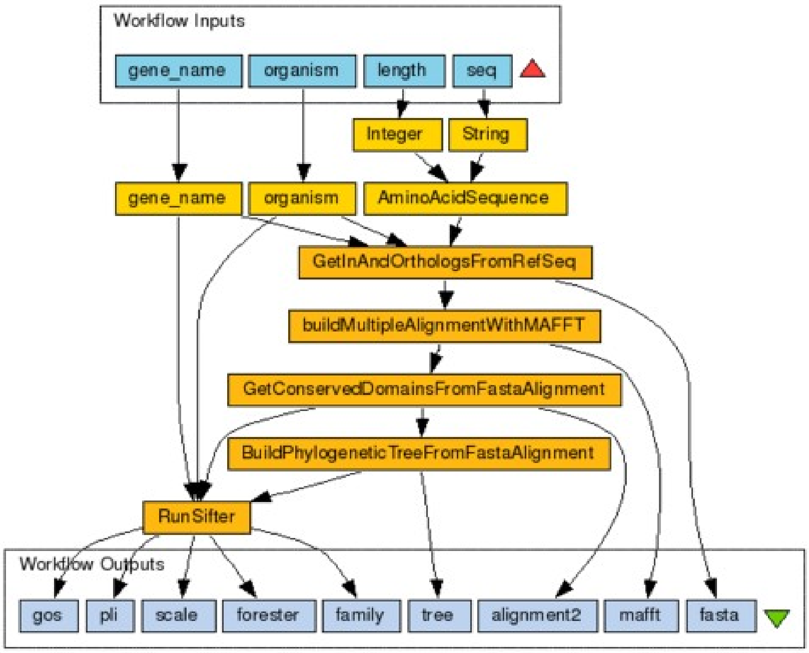


Figure 8: Taverna SIFTER workflow in MyExperiment.

c) GO term annotation

Medicago genome project

The 20060904_imgag_protNONRED.fa fasta file, which includes protein sequences of all predicted Medicago genes in the first release of the Medicago genome (MT1.0), was downloaded from a secured website at the MIPS institute in Munich[12]. Each Medicago protein was extracted from the fasta file, stored in a separate fasta file and used as input for the Medicago SIFTER pipeline.

InterProScan was run at the J. Craig Venture Institute in Washington DC. Because there was a problem with many false positives for the PFAM family [Bateman et al. 2002] prediction, the PFAM calculation was run again at the MIPS institute by Thomas Rattei. The former PFAM family analysis was updated by the new analysis results. InterPro accessions were mapped via the InterPro2GO file to molecular function GO terms. The InterPro2GO mapping file was downloaded from the Gene Ontology website[13] on 5 September 2006. Mapped GO terms were afterwards combined with predicted GO terms by the Medicago SIFTER pipeline.

Sorghum genome project

Protein sequences of the 27458 predicted Sorghum proteins were kindly provided as a fasta file by the MIPS institute in Munich, which is part of the Sorghum genome project. This file was split in 54 different fasta files to run the SIFTER pipeline iteratively in parallel with each Sorghum protein as input. All results were stored in the AFAWE MySQL database to enable a fast evaluation of them.

InterProScan in combination with InterPro2GO was not used for the GO term annotation.

Tomato genome project

The amino acid sequence of the 9942 genes available in batch11 were downloaded on 27 May 2008 from the SGN sFTP server at upload.sgn.cornell.edu. The improved phylogenomic pipeline with SIFTER (see chapter V2b) was run iteratively using the amino acid sequence of each gene as input.

InterProScan was run at the Imperial College in London and the results were uploaded to the SGN sFTP server. I downloaded the InterProScan results on 6. June 2008 from the SGN sFTP server and extracted all InterPro accessions predicted to the corresponding genes via a Perl script. Afterwards InterPro accessions were mapped via the InterPro2GO file, which was downloaded from the Gene Ontology website on 9. May 2008, to molecular function as well as biological process GO terms.

Molecular function GO terms predicted by the SIFTER pipeline and by InterProScan in combination with InterPro2GO were combined by a Perl script, written to a separate file and uploaded again to the SGN sFTP server.

d) Comparison of the number of proteins annotated in the most general GO term categories

To get an overview of the number of proteins in the main Gene Ontology categories and to compare them with GO term annotated genomes from other species, a Perl script was implemented, which gets for each assigned GO term the second level GO term parent (GO categories below GO term "molecular function") and counts the number of proteins for these categories. The same script was also used to count the number of proteins for each of the main "molecular function" GO categories

[12] *http://mips.gsf.de/proj/medicago/secure/20060904_imgag_protNONRED.fa*
[13] *http://www.geneontology.org/external2go/interpro2go*

in the *Arabidopsis thaliana, Oryza sativa, Rattus norvegicus, Homo sapiens* and *Mus musculus* genome. GO term assignments for these species were extracted from annotation files provided from different institutes via the Gene Ontology website.

e) The accuracy of SIFTER

To determine the false discovery rate of the SIFTER workflow, the predictions of 100 Medicago proteins and 100 Sorghum proteins were manually checked. Stephan Schlößer, a practical student from the University of Cologne, helped with the manual annotation of the Sorghum proteins.

Blast searches against different databases (RefSeq [Pruitt et al. 2008] [Sayers et al. 2008], UniProt [UniProt 2007], nr [Sayers et al. 2008] and TAIR [Weems et al. 2004]) were run using an e-value cutoff of 1. Additionally InterProScan and RPSBlast against the Conserved Domain Database (CDD) [Marchler-Bauer et al. 2007] were run to search for conserved protein domains in the Medicago proteins. GO terms, EC numbers and description lines from all Blast hits, InterProScan and RPSBlast results were compared with the predicted GO term by SIFTER. It is assumed that the predicted function of the corresponding Medicago protein is true, if the following criteria are true:

- The Medicago protein has a Blast hit in one of the databases with the following criteria
  - overlap between query and hit > 80%
  - functional description line of the hit protein is semantically the same as the predicted GO term or includes the predicted function and includes no “putative”, “like”, ”probable” or “similar to”
  - predicted GO term is experimentally proven or reviewed for the hit protein
  - query and hit protein have the same protein domains
  - predicted function is shown to be true in the literature for the hit protein
- In InterProScan or RPSBlast results the Medicago protein has a protein domain, which has the same GO term as predicted by SIFTER or the predicted function is described in the functional description for one protein domain
- predicted function of the protein is published in the literature

Manually checked functions, which are confirmed based on these criteria are annotated to the corresponding Medicago proteins.

f) Looking for genes, which are unknown in plants, but have been functionally characterized in animals

To find genes in the Medicago and the Sorghum genome which have homologous proteins in animals but not in plants a Blastp search using all Medicago proteins and all Sorghum proteins as query against the UniProtKB database (release 9.0) was run. All Blast results were stored in the AFAWE MySQL (see figure 22) database and those which have no hits to model plants like *Arabidopsis thaliana* and *Oryza sativa*, but a significant hit ($< e^{-5}$) to any other non-plant-organism were automatically extracted by a Perl script. These candidate genes were further investigated manually by using the following criteria:

- query has no hit in plants in any other database (nucleotide database or protein database) using different Blast programs

- found domains by InterProScan and RPSBlast in the query are also included in the homologous genes in other organisms
- good sequence conservation between query and hit

Interesting genes were classified in gene families, if more than one member was found. Additionally, for the Transferrin family, found in the Medicago genome, the evolutionary history was explored by investigation of a bootstrap phylogenetic tree for one conserved domain to increase the accuracy. All amino acid sequences were aligned using MAFFT [Katoh et al. 2005] (Parameters: --auto), transferred by a Perl script to the PHYLIP format and the alignment was used as input for seqboot from the PHYLIP package [Felsenstein 1993] to generate 100 bootstrap sequences. The bootstrapped sequences were used as input for proml [Felsenstein 1993] to build maximum likelihood trees and these were combined using consense [Felsenstein 1993]. Homologous genes were extracted from published papers and from Blast results (The Blast search tool from the NCBI website [Johnson et al. 2008] was used) against the UniProt database [UniProt 2007] and against EST databases [Sayers 2008]. Domain positions were extracted by using InterProScan [Mulder and Apweiler 2007] and the amino acid sequence for the first domain was extracted by using extractseq from the EMBOSS package [Rice et al. 2000].

g) Verification of the gene prediction results in the tomato genome project

To verify the predicted gene structure on which the functional annotation relies and to explain functional prediction results, Blastp results of all tomato genes from batch11 against the TAIR7 database [Weems et al. 2004], computed by members from the inter-disciplinary Centre for Plant Genomics in the department of Plant Molecular Biology at the University of Delhi, India, were downloaded from the SGN sFTP server. Tomato genes were counted, for which the overlap between query and best hit was below 50%, 60% and 70% using a self written Java program. The overlap was computed by using equation 1 (see chapter V2a). Some potentially wrongly annotated genes were manually inspected afterwards.

3) Results

a) GO term annotation

GO term annotation in the Medicago genome project

13978 Medicago proteins could be assigned at least one GO term by using the Medicago SIFTER pipeline and InterProScan in combination with InterPro2GO. 4853 proteins got GO term assignments only from InterProScan, 2183 proteins only from SIFTER and 6911 proteins from both tools (see figure 9).

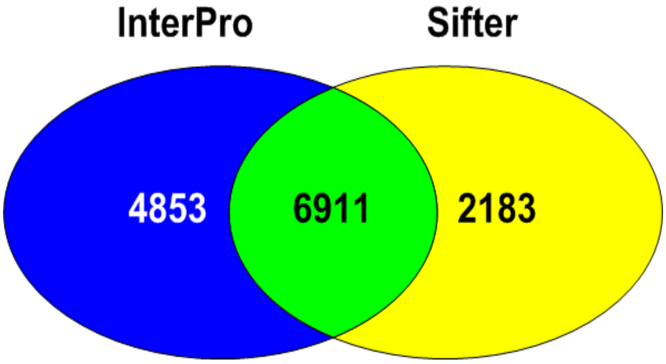


Figure 9: Comparison between InterProScan and SIFTER in number of annotated genes.

A comparison of the number of annotated proteins in the main “molecular function” Gene Ontology categories from *A. thaliana*, *O. sativa* and *M. truncatula* revealed no significant conspicuities (see figure 10). Triplet codon amino acid adapter activity was only assigned by *A. thaliana*, because tRNA genes were not included in the *M. truncatula* and *O. sativa* genome. There seem to be a difference between genes involved in binding processes between *A. thaliana, M. truncatula* and *O. sativa.*

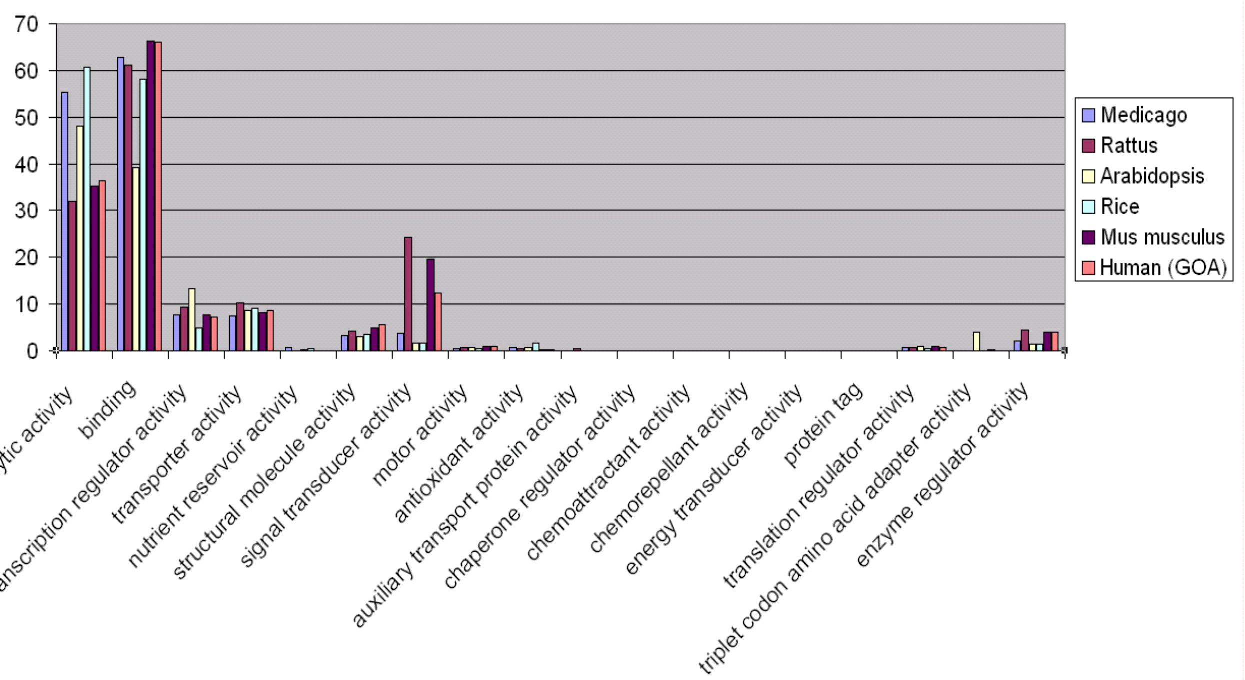


Figure 10: Comparison between genomes of different species and the Medicago genome in number of annotated genes in the most general "molecular function" Gene Ontology categories. The number of genes (Y axis) is expressed as a percentage.

By comparing the number of annotated genes in plant genomes with the number of annotated genes in animal genomes I found three significant differences. One of them is the number of genes annotated as signal transducer activity (see figure 10 and table 4). There are more genes in animals annotated with GO term signal transducer activity than in any plant species. In Medicago 257 proteins of the 512 annotated proteins in the category “signal transducer activity” got the GO term GO:0004888 (transmembrane receptor activity) assigned. 153 of the 257 have a coiled coil (CC), a nucleotide binding domain and a leucine rich repeat (LRR) domain and so are probably members of the family of disease resistance proteins. In 97 cases only the TIR domain was present. All genes show similarity to the Mal/TIRAP genes in human, which are not receptors. Mal/TIRAP genes are involved in the Toll-like receptor signaling pathway which is used in innate and adaptive immunity and are ubiquitously expressed in the cytoplasm [Brikos and O'Neill 2008] [Martin and Wesche 2002] [Horng et al. 2001]. In Arabidopsis the number of disease resistance proteins and proteins which have only the TIR domain is 142. Another difference in the GO term analysis is the number of annotated genes with the GO term “nutrient reservoir activity” (see table 4). The number of proteins annotated with that GO term seems to be increased in plants. The last difference concerns the GO category “enzyme regulator activity”. Here again the number of animal genes is slightly increased.

Because the annotation file from each organism is submitted from different institutes, the number of genes with experimentally verified or reviewed GO terms were completely different for each organism. The annotation files for the model organism *A. thaliana* and *H. sapiens* included the most experimentally verified or reviewed GO terms. By contrast the number of annotated genes in the *R. norvegicus* genome was very low (see table 4).

GO term annotation in the Sorghum genome project

Using the new pipeline we were able to assign 15123 of the 27458 (~55%) predicted Sorghum protein coding genes up to three different GO terms. In contrast to the former pipeline, which took about three months for all 43616 Medicago genes to be finished, calculating all 27458 Sorghum genes, using the new pipeline, just took approximately two weeks by using 10 computenodes (dual processor) of the compute cluster.

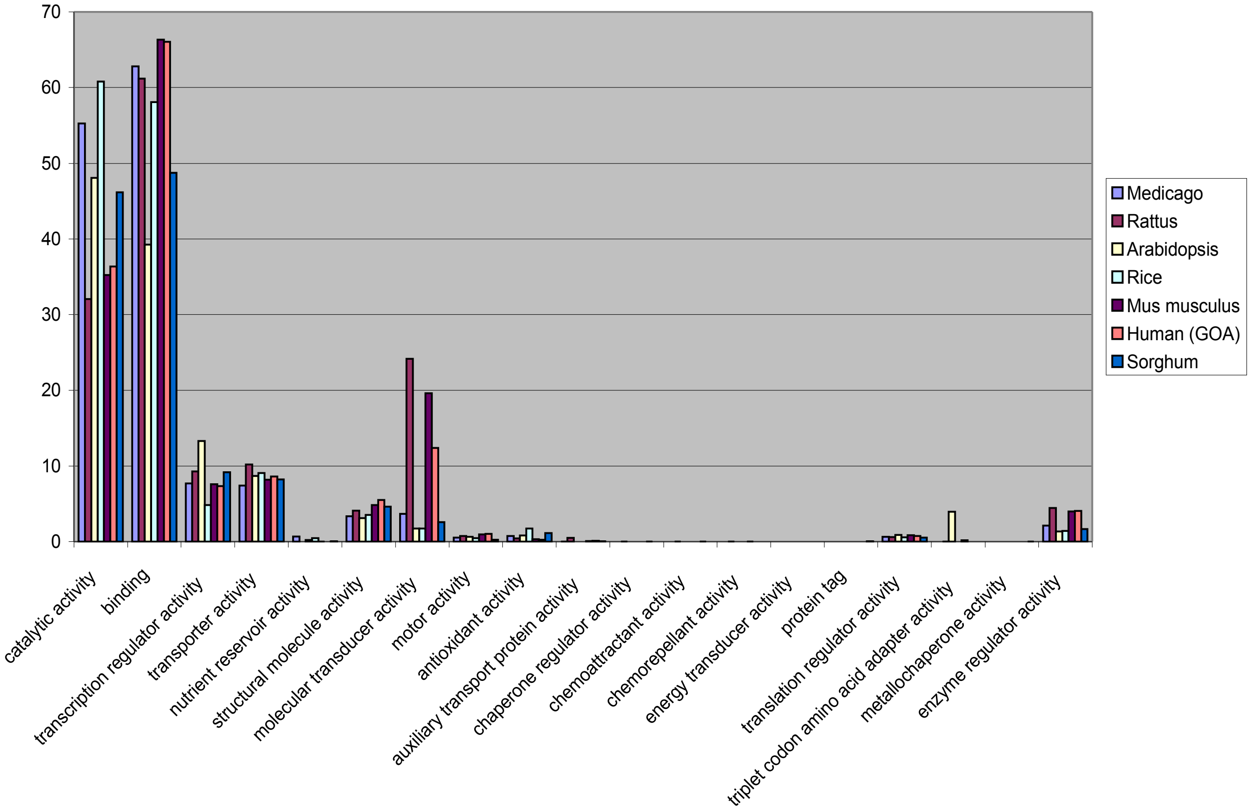


Figure 11: Comparison of the number of genes in the most general "molecular function" Gene Ontology categories between different organisms.

The comparison between the number of genes in the second level “molecular function” Gene Ontology categories in Sorghum, Arabidopsis, rice, mouse, rat and human revealed no conspicuities in the Sorghum genome. However, it is remarkable that the number of Sorghum genes in GO category “transcription regulator activity” is about two times higher than the number of genes in the Sorghum closest relative *Oryza sativa* (see table 4). Also in the categories “molecular transducer activity” (new parent of “signal transducer activity” in the GO graph) and in “structural molecule activity” there are more Sorghum genes annotated with that category than in rice. In contrast the number of rice genes in category “nutrient reservoir activity” is approximately six times higher than in Sorghum. And also in category “motor activity” the number of rice genes in higher than in Sorghum.

GO category	***Medicago truncatula***	***Sorghum bicolor***	***Oryza sativa***	***Arabidopsis thaliana***	***Mus musculus***	***Humo sapiens***	***Rattus novegicus***
Number of annotated genes	13947	15123	13330	17396	16187	29030	9931
catalytic activity	7708	6978	8102	8362	5703	10559	3184
binding	8759	7370	7741	6828	10734	19177	6077
transcription regulator activity	1072	1388	648	2314	1231	2127	923
transporter activity	1032	1244	1206	1511	1326	2494	1014
nutrient reservoir activity	96	11	60	37	0	0	0
structural molecule activity	466	702	470	539	780	1607	407
molecular transducer activity	512	390	230	299	3176	3597	2400
motor activity	75	37	66	109	157	301	76
antioxidant activity	104	175	231	142	56	76	43
auxiliary transport protein activity	4	7	1	1	13	33	51
chaperone regulator activity	1	1	0	0	1	6	4
chemoattractant activity	0	0	0	0	1	5	4
chemorepellant activity	0	0	0	0	4	0	3
energy transducer activity	0	0	0	0	0	0	0
protein tag	1	10	0	0	1	0	0
translation regulator activity	91	84	75	159	138	215	60
triplet codon amino acid adapter activity	0	0	0	688	28	0	2
metallochaperone activity	0	3	0	0	0	0	0
enzyme regulator activity	297	252	188	236	650	1182	443

Table 4: Number of genes in the most general "molecular function" GO categories of M. truncatula, S. bicolor, O. sativa, A. thaliana, M. musculus, H. sapiens and R. norvegicus

GO term annotation in the Tomato genome

1915 tomato genes (~19%) were annotated with up to three molecular function GO terms by using the improved SIFTER pipeline. In comparison to that, InterProScan in combination with InterPro2GO annotated molecular function GO terms to 3050 proteins (~31%). By combining the results from both analyses 3478 genes (~36%) could be functionally annotated with molecular function GO terms. 2102 (~21%) tomato genes were annotated with at least one biological process GO term.

GO category molecular function	*Number of Tomato genes*
Catalytic activity	1889
Binding	2401
Transcription regulator activity	240
Transporter activity	252
Nutrient reservoir activity	9
Structural molecule activity	80
Molecular transducer activity	78
Motor activity	20
Antioxidant activity	46
Auxiliary transport protein activity	2
Chaperone regulator activity	1
Protein tag	1
Translation regulator activity	8
Enzyme regulator activity	55

Table 5: Number of genes annotated in the most common molecular function Gene Ontology categories by the phylogenomic pipeline with SIFTER and InterProScan in combination with InterPro2GO.

The number of genes in the most general molecular function GO categories is shown in table 5 and the number of genes in the common biological process GO categories is shown in table 6.

In most biological process GO categories at least one gene is present, except for the categories “Cell killing”, “Growth”, “Pigmentation” and “Rhythmic process”.

GO category biological process	*Number of Tomato genes*
Biological adhesion	1
Biological regulation	304
Cell killing	0
Cellular process	1521
Development process	51
Establishment of localization	297
Growth	0
Immune system process	7
Localization	300
Metabolic process	1626
Multi-organism process	13
Multicellular organismal process	6
Negative Regulation of biological process	2
Pigmentation	0
Positive Regulation of biological process	1

GO category biological process	*Number of Tomato genes*
Regulation of biological process	255
Reproduction	9
Reproductive process	9
Response to stimulus	139
Rhythmic process	0
Viral reproduction	6

Table 6: Number of genes annotated in the most common biological process Gene Ontology categories by InterProScan in combination with InterPro2GO.

The distribution of the number of genes in the most general molecular function GO categories is approximately the same as for other plant genomes (see figure 12).

To 6329 genes no GO term (molecular function as well as biological process) could be annotated.

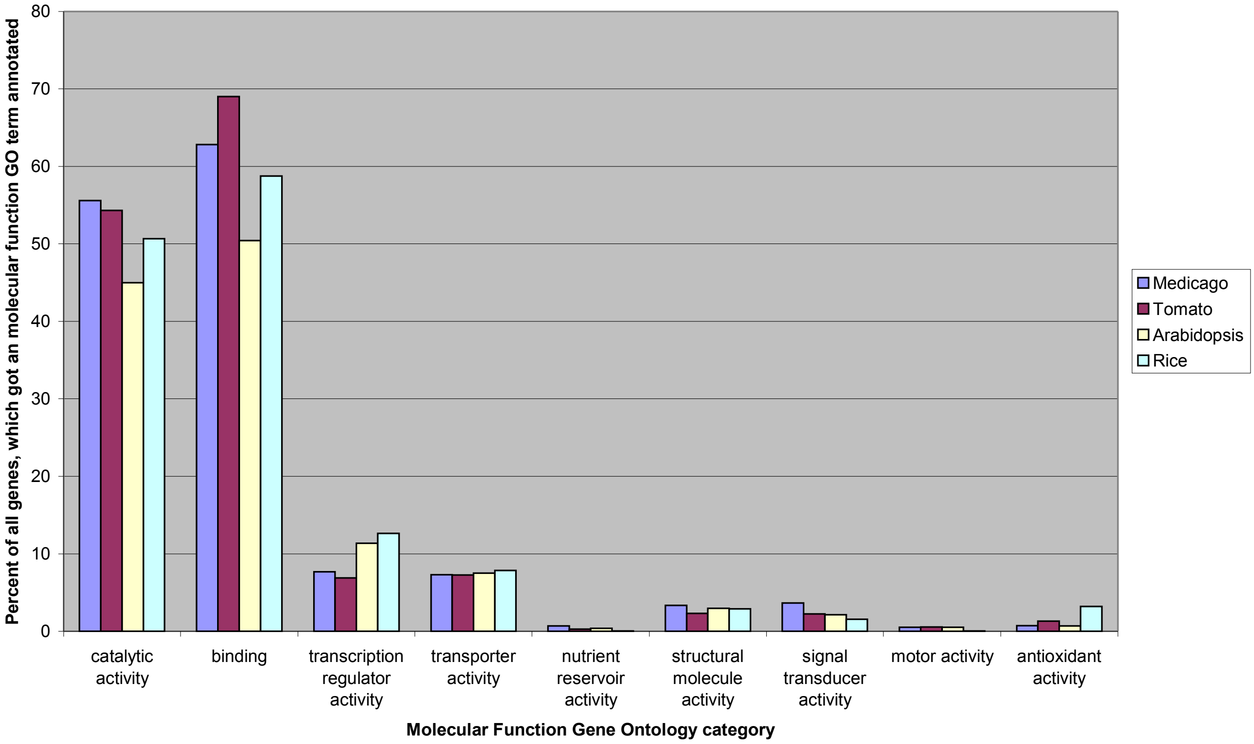


Figure 12: Comparison between the number of genes in the most general Gene Ontology categories from different plants.

b) SIFTER accuracy

Accuracy of SIFTER using the Medicago pipeline

I could identify three Medicago genes with wrong annotations made by SIFTER (see table 7).

Gene identifier	*SIFTER annotation*	*Manual annotation*
AC144389_35.2	cytochrome-c oxidase activity	Cytochrome b5
AC149803_8.2	flavonol synthase activity	ACC oxidase
AC149131_5.2	retinal dehydrogenase activity	NAD dependant epimerase

Table 7: Wrong annotations made by SIFTER.

AC144389_35.2 was assigned a wrong GO term, because a homologous protein from human has a wrong GO term annotation (cytochrome-c oxidase activity instead of Cytochrome b5) with evidence code TAS (Traceable author statement). In case of AC149803_8.2 and AC149131_5.2 GO term "flavonol synthase activity" and GO term "animal retinal dehydrogenase" assigned to proteins inside the phylogenetic tree were experimentally verified or published in a paper and therefore got a higher probability than GO term "3-ketoacyl-(acyl-carrier-protein) reductase" (See figure 13) and GO term "oxidoreductase activity" which are only reviewed (evidence code ISS) and are annotated to the NAD dependant epimerase proteins. Furthermore the GO term "oxidoreductase activity" is the parent of the "animal retinal dehydrogenase" GO term and is therefore not considered as

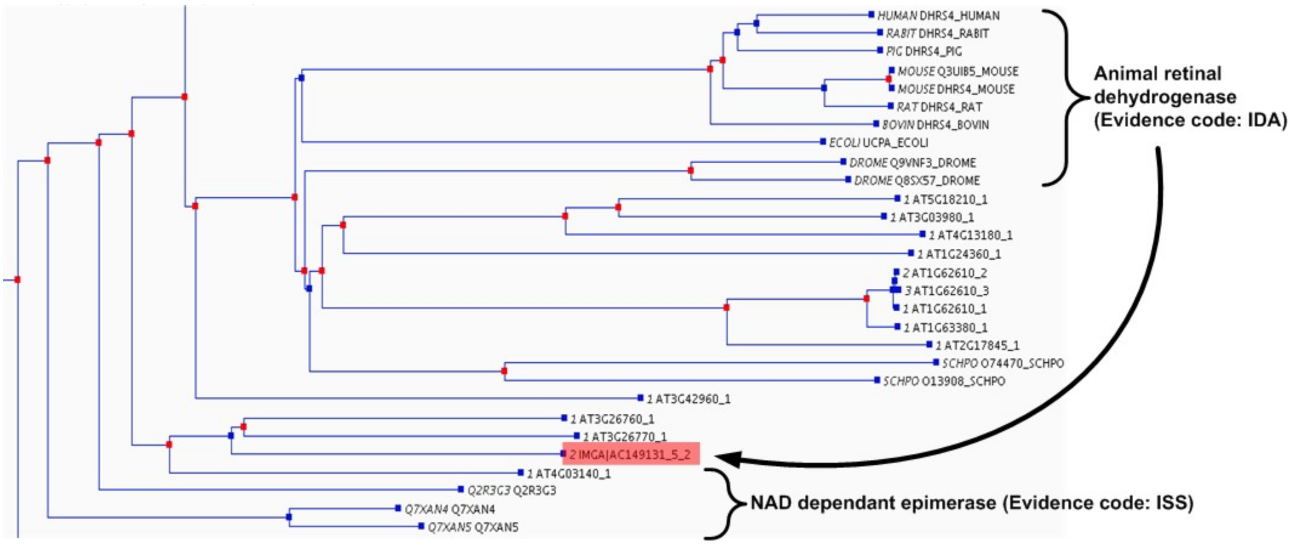


Figure 13: Wrong prediction made by SIFTER in case of Medicago gene AC149131_5.2.

Another problem is, that SIFTER assigns only one GO term for each protein, so many proteins lack GO terms. However, SIFTER was in 25% more specific than the assigned human readable description and achieved an accuracy of 97%, which was better than running a simple Blast search (94% accuracy). InterProScan assigned in all cases right GO terms, but these were very general.

Accuracy of SIFTER using the improved pipeline

By using the improved SIFTER pipeline for the GO term annotation of the 100 tested Medicago genes, the prediction accuracy could be increased to 100%.

However, by comparing the assigned manual annotation of 100 Sorghum protein coding genes with the predicated GO term by SIFTER, GO terms with the best probability of seven Sorghum genes were wrong (see table 8). Three of them are structure proteins and signaling proteins, which have many low complexity regions in their sequence and therefore many wrong homologs have been found in the first step of the pipeline. Another problem was, that a homologous gene of one Sorghum protein has a wrong annotation assigned by traceable author statement. This wrong annotation led to a wrong annotation of the Sorghum protein. Further problems were missing GO terms and the transfer of GO terms which were experimentally verified or reviewed by an author, to all leaves (proteins) in the whole phylogenetic tree.

The GO terms of 18 Sorghum proteins with the lowest probability calculated by SIFTER were also wrong. But except in one case, the probability of the wrong assigned GO terms was lower than 0.4.

Sorghum identifier	*Predicted GO term*	*Manual annotation*
Sb01g032650	extracellular matrix constituent conferring elasticity	Copper transport protein ATOX1-related.
Sb04g020710	retinol dehydrogenase activity	11BETA-Hydroxysteroid dehydrogenase
Sb03g017580	NADH dehydrogenase activity	Cytochrome b
Sb01g038400	citrate transporter activity	Mitochondrial succinate-fumarate transporter
Sb10g029800	extracellular matrix constituent conferring elasticity	Unknown transcription factor with heterodimerization activity
Sb10g027460	GABA-A receptor activity	Plastocyanin-like domain containing protein
Sb04g023830	protein homodimerization activity	aldose 1-epimerase

Table 8: Sorghum genes, for which the best predicted GO term is wrong.

c) Non-plant specific genes

Non-plant specific genes in Medicago

I could identify four proteins in the Medicago genome with no significant hit to Arabidopsis and rice, but hits to animals like human, mouse and rat. One of them (AC174375_1.1) is a putative DNA topoisomerase. Because the best hits are to genes of Plasmodium, this protein could be a horizontal gene transfer from protozoa.

The other three proteins (AC126015_39.1, AC153430_18.1 and AC153430_10.1) belong to the transferrin protein family, which is not known in higher plants yet. Members of the transferrin family are widely distributed in all kinds of organisms (except fungi). They are glycosylated proteins that transport iron from plasma to cells or help regulate iron levels in biological fluids. Many different subfamilies are known [Lambert et al. 2005]. They all share the same iron binding transferrin domain, but most of them have two or three transferrin domains that resulted from a gene duplication event occurring around 850 Mya [Park et al. 1985]. Besides the well known serum transferrin in animals members of that family are also found in insects [Huebers et al. 1988] [Kurama et al.1995] [Thompson et al. 2003], sea urchins (mayor yolk proteins (MyP)) [Brooks and Wessel 2002] and in Duniella algae [Fisher et al. 1997] [Fisher et al. 1998].

One of the three Medicago transferrins (AC153430_18.1) was excluded from further analysis, because almost the whole transferrin domain was not yet sequenced. The other two transferrin like genes have one complete, well conserved transferrin domain. By comparing their protein sequences it was found that these genes are 100% identical and so it is assumed that they are actually the same gene and are redundant in our input dataset. Both genes are located on two different BACs.

After running Blast searches against different EST databases (see chapter V2f) further members of the transferrin family in plants were detected in *Citrus clementina*, *Picea glauca*, *Picea sitchensis*, *Amborella trichopoda*, *Adiantum capillus-veneris* and *Pseudotsuga menziesi.* Transferrin family members were also discovered in the cyanobacteria *Anabaena variabilis* and *Nostoc sp.*.

Lambert et al. published a comprehensive phylogenetic tree with 71 transferrin family sequences from 51 species [Lambert et al. 2005]. They assumed that the transferrin members in algae may represent a horizontal gene transfer event, because they clustered quite well with transferrin-like sequences from insects. To find out where in the phylogenetic tree the all plant transferrin proteins fit, all available sequences from Lamberts et al. were used together with the Medicago proteins to build the tree again. To increase the accuracy of the tree, only the first transferrin domain from each protein was used (for material and methods see chapter V2f).

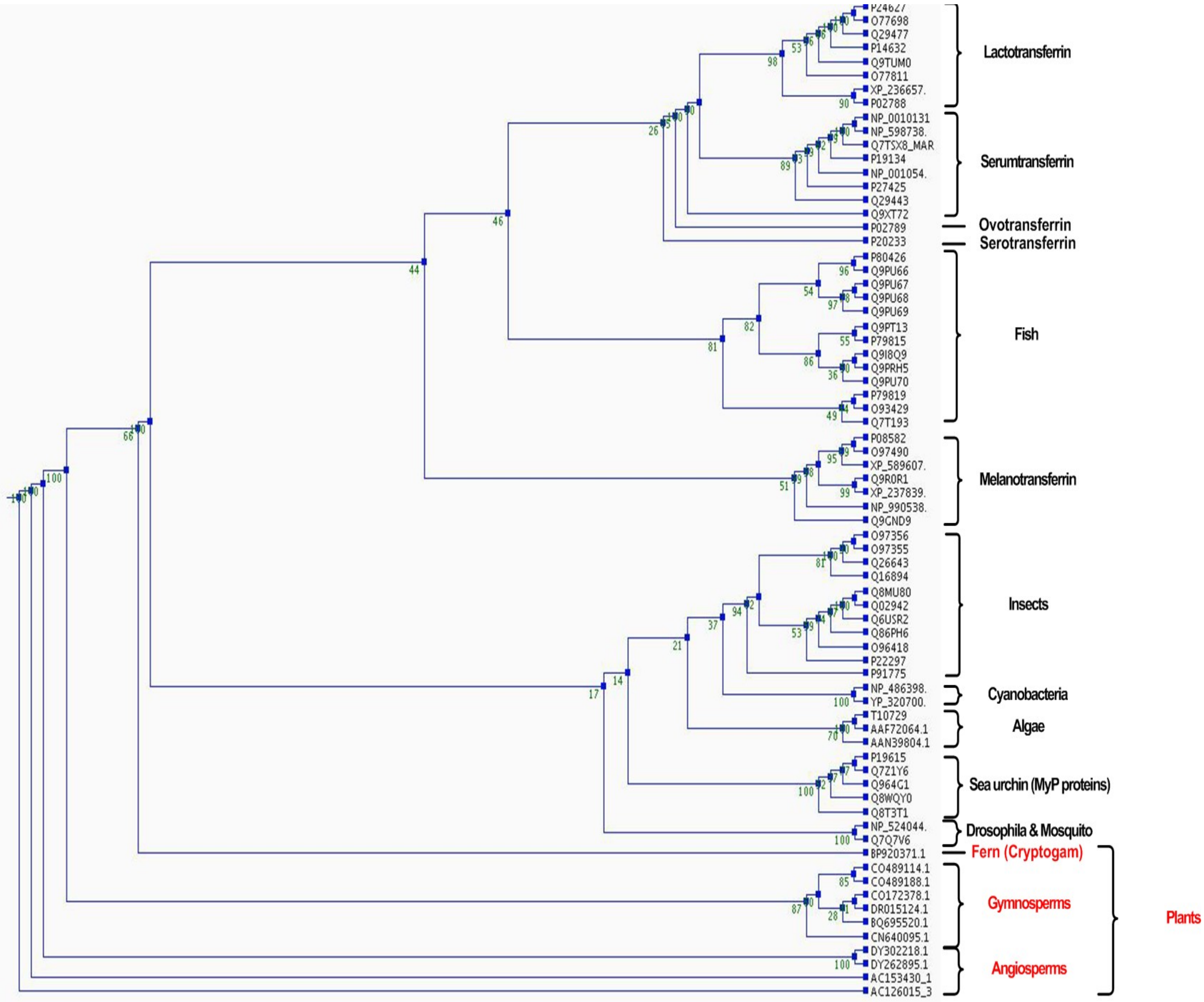


Figure 14: Phylogenetic tree of the first transferrin domain of proteins from Lambert et al. and transferrin proteins found in plants and cyanobacteria. Please note that this is an unrooted tree and has no branch lengths.

As shown in figure 14 all plant transferrins, including the Medicago proteins (AC153430_1 & AC126015_3), are in the same subtree as the MyP proteins from sea urchins, proteins from insects, proteins from algae and cyanobacteria transferrins. Plant transferrins split into three subgroups. One subgroup includes transferrins of all Picea species and *Pseudotsuga menziesii*, these organisms belong to the Gymnosperms [Palmer et al. 2004] [Farjon et al. 1991]. The second subgroup contains transferrins from *Citrus clementina* and *Medicago truncatula*, which belong to the Angiosperms. The third subgroup includes only the transferrin protein from the fern *Adiantum capillus-veneris*. The fern transferrin-like genes seem to be the evolutionary oldest genes in comparison to Angiosperm and Gymnosperm transferrin-like genes, because they are in the phylogentic tree close to algal and cyanobacterial transferrins. Angiosperm transferrin genes are in comparison to fern and

Gymnosperm transferrin-like genes the most evolved genes (see figure 14). This fits well with the plant tree of life described in [Palmer et al. 2004].

The whole plant, insect, sea urchin, insect, algae and cyanobacteria transferrin subfamily looks most related to Melanotransferrins, which are assumed to be the oldest vertebrate transferrins [Baldwin et al. 1993].

The phylogenetic tree of plant, insects, algae, cyanobacteria and sea urchins reflects the evolutionary history from primitive old organisms (e.g. algae, cyanobacteria or ferns) to higher evolved, younger organisms like Angiosperms and insects.

Non-plant specific genes in Sorghum

14 Sorghum genes were defined as "non-plant-specific", because they show a high similarity with genes in non-plant organisms (see table 9). Three of them (Sb07g002440, Sb06g031470 and Sb06g019820) could be horizontal gene transfers from bacteria, because besides hits to rice, moss (*Physcomitrella patens*) and the common grape vine (*Vitis vinifera*), the best Blast hits are from bacteria. Sorghum genes Sb03g04600 (possible F-Box protein), Sb08g005120 and Sb01g048230 (calcium binding protein) had hits in plants, but not in Arabidopsis. In contrast genes Sb04g004130, Sb06g028290 and Sb05g006000 do not have any hits in plants, but share similarity with animal genes and Drosophila genes. One gene (Sb01g003920) could be a transposon and the other four genes seem to be annotation errors, because in an alignment with Arabidopsis, genes with the same assigned function aligned in some regions very well.

Sorghum identifier	***Hits in plants***	***Horizontal gene transfer?***	***Manual functional annotation***	***Putative annotation error?***
Sb07g002440	Yes, in *Oryza sativa,Vitis vinifera* and *Physcomitrella patens*	yes	ATP-DEPENDENT CLP PROTEASE	no
Sb06g031470	Yes, in *Oryza sativa, Vitis vinifera* and *Physcomitrella patens*	yes	unknown protein Cupin 4	no
Sb06g019820	Yes, in *Oryza sativa* and *Vitis vinifera*	yes	Beta-ketoacyl synthase	no
Sb03g043600	Yes, but no hits to Arabidopsis genes	no	Putative F-Box	no
Sb08g005120	Yes, in *Oryza sativa* and *Vitis vinifera*	no	Similarity to human Mature T-cell proliferation	no
Sb01g048230	Yes in *Vitis vinifera*, Populus and *Picea sitchensis*	maybe	Calcium binding protein with two EF-hands	no
Sb01g003920	Yes, but low significance hits to Arabidopsis.	no	Putative transposon	no

Sorghum identifier	***Hits in plants***	***Horizontal gene transfer?***	***Manual functional annotation***	***Putative annotation error?***
	Better hits to *Oryza sativa*, *Vitis vinifera*, *Olimarabidopsis pumila* and plasmodium			
Sb04g004130	no	no	Shows low similarity to human receptor genes and InterProScan predicts a signal peptide	no
Sb06g028290	no	no	low similarity to human Pre-mRNA-splicing factor 38B	no
Sb05g006000	no	no	Similarity to drosophila Papilin precursor	no
Sb01g026250	Yes	no	UBIQUITIN poly protein	yes
Sb06g017200	Yes, but no hits to Arabidopsis genes	no	DNA-DIRECTED RNA POLYMERASE	yes
Sb06g021110	Yes, low significance hits to rice and Arabidopsis	no	Similarity to drosophila and "Skin secretory protein xP2 precursor protein" of Xenopus laevis	yes
Sb08g013610	no	no	ADP ribosylation factor	yes

Table 9: Non-plant specific genes found in the Sorghum genome

d) Validation of the gene prediction results in the Tomato genome project

To check if the reason for the low number of annotated Tomato genes by SIFTER is poor gene prediction, a Blast search against the Arabidopsis genome was evaluated (see Material & Methods section). By checking the overall overlap between query and hit approximately 30% of all tomato genes have an overlap smaller than 60% between the amino acid sequence of the tomato gene and the best Blast hit (see table 10). Furthermore 2785 tomato genes (28%) have an overlap between query and hit sequence below 60% at an identity greater than 30% and an expectation value below e-5. However, by comparing the frequency of alignments where the tomato gene has the greater overlap compared to the Arabidopsis hit, it can be assumed that many tomato genes are truncated.

Criterion	***Number of genes***
Overlap between query and best Blast hit < 50%	2670 (~27%)
Overlap between query and best Blast hit < 60%	3005 (~30%)
Overlap between query and best Blast hit < 70%	3327 (~33%)

Table 10: Number of tomato genes with an overall overlap smaller than 50%, 60% and 70% to the best matching Arabidopsis gene.

Criterion	*Number of genes*
Query overlap < 50%	505 (~5%)
Query overlap < 60%	836 (~8%)
Query overlap < 70%	1284 (~13%)

Table 11: Overlap between aligned region and query is smaller than the threshold, i.e. the tomato sequence is longer.

Criterion	*Number of genes*
Hit overlap < 50%	2580 (~26%)
Hit overlap < 60%	2908 (~29%)
Hit overlap < 70%	3213 (~32%)

Table 12: Overlap between aligned region and hit is smaller than the threshold, i.e. the tomato sequence is shorter.

4) Discussion

GO term annotation

I provided two automatic pipelines suitable for an accurate and fast automatic functional annotation in genome projects. Both pipelines were tested on the first release of the Medicago genome and the improved pipeline was further tested on the finished Sorghum genome and in the on-going tomato genome project. In comparison to the pipeline used in the Medicago genome project, the improved pipeline is reusable in other genome projects and can be easily modified.

Using the Medicago SIFTER pipeline in combination with InterProScan and InterPro2GO ~32% of the available Medicago proteins (~60% of the genome) could be annotated with GO terms. By using more than one functional analysis tool the number of annotated genes could be increased and the function of a protein could be further specified. The Medicago SIFTER pipeline was only able to annotate approximately 21% of all predicted Medicago genes. One possible reason for that could be the erroneous structural annotation of the Medicago genome and the not-finished assembly of the genomic sequence. A hint to that could be that many Medicago proteins (~17000) do not have a hit in any database with an overlap greater than 70% and the exon/intron boundaries and the open reading frame of many of them seem to be wrongly predicted by the gene finders. This problem could be solved by further training of the gene finders with manually curated gene models from *M. truncatula*. This manual curation effort could be done in a community approach [Thibaud-Nissen et al. 2007]. Another reason could also be that many genes are not detected in the genomic sequence of Sorghum (Underprediction). Maybe this result could be further improved by using more sensitive methods for the remote homolog detection like Hidden Markov Models (HMMs) [Durbin et al. 1998] (e.g. FlowerPower [Krishnamurthy et al. 2007]) or a profile search as implemented in PsiBlast [Altschul and Koonin 1998] and FastBlast [Price et al. 2008] using the query sequence as a template. However, this may leads (e.g. in case of HMMs) to a speed reduction of the whole pipeline.

The improved SIFTER pipeline was much faster. One time-consuming step in both pipelines was the tree building process. To further speed up the pipeline, methods [Wapinski et al. 2007] which incorporate the query sequence in pre-built trees extracted from phylogenetic tree databases like PhyloFacts [Glanville et al. 2007] and TreeFam [Li et al. 2006] [Ruan et al. 2008] should be used. Unfortunately such databases are as yet rare for plants.

In batch11 of the tomato genome project I was able to annotate 3478 (~35%) of the predicted 9942

tomato genes with at least one molecular function GO term and 2102 tomato genes (~21%) with at least one biological process GO term.
The improved SIFTER pipeline was able to annotate only 1915 tomato genes. InterProScan in comparison was able to annotate many more tomato genes (3050 (~31%)). This could have several reason. One explanation could be that the gene prediction is poor. A hint to that could be that approximately 30% of all genes have an overlap of less than 60% with the best Blast hit. 28% of these have a sequence identity greater than 30% and an expectation value below e-5. The overlap was used in the pipeline to separate homologous genes from non-homologs and it seems that many homologous genes could not be detected. In this case InterProScan has the advantage that it is able to annotate also incomplete or over-predicted genes, because it just looks for protein domains or motifs in the sequence and does not consider the whole gene structure. Furthermore some tools (e.g. Hidden Markov Model search) integrated in InterProScan are more sensitive than Blast. I assume that many gene predictions are too short or split into two genes by the gene finder, because the alignment covers > 60% of the query, but below 60% of the hit. Also the tomato genes I checked manually confirmed that. One way to deal with this in the SIFTER pipeline could be to lower the cutoff. However, this would introduce many non-homologous genes and would lead to false annotations.

Another reason for the low number of annotated tomato genes could be that many tomato genes are shorter than the homologous Arabidopsis genes. However, there is no evidence that this could be true. But maybe many tomato genes are not included in the Arabidopsis genome and therefore the best Blast hit is not even a homologous gene. In this case also the overlap between the tomato sequence and the Arabidopsis sequence would in most cases be low and this is not the case in the evaluation of the Blast result and most tomato genes show a high sequence conservation to the best matching Arabidopsis gene.

SIFTER accuracy

By using the Medicago SIFTER pipeline the false discovery rate for 100 manually annotated Medicago proteins was only 3% and SIFTER was in 25 cases more specific than the assigned human readable annotation. The false discovery rate can be further increased to 100% by using the improved SIFTER pipeline. The difference between the results can be explained by the more comprehensive phylogenetic tree provided by the improved SIFTER pipeline. By looking only for putative homologs in a database with genes, which have an experimentally verified or reviewed GO term, the final phylogenetic tree is in many cases incomplete. Several in-paralogous genes in Medicago may also be missing, because the sequencing progress of Medicago is still in progress (status 2007). So forester assigned in several trees wrong duplication and speciation nodes, which can lead to wrong annotations. But these results indicate that SIFTER is able to assign specific GO terms to genes with a low false discovery rate.

But also using the improved SIFTER pipeline, 18 Sorghum proteins of the Sorghum test set got wrong GO term annotations. The number of wrong annotations can be decreased to seven by applying a posterior probability cutoff of 0.4. In addition wrong GO term annotations for three of the seven Sorghum proteins could have been avoided by using the low complexity filter of Blast (parameter: -F T) for the iterative Blast search. This parameter was integrated afterwards in the pipeline to improve further predictions.

However, I detected several problems using SIFTER. SIFTER was very slow in version 0.3 especially for huge gene families. If there were only few GO terms with low evidence level available for proteins inside the phylogenetic tree SIFTER tends to wrong annotations. Another problem was, that SIFTER assigned only the GO term with the highest probability. This is the default behavior in SIFTER, but can be changed by editing the source code.

Comparison between the number of genes annotated in the most general molecular function GO categories

A comparison of the percent of genes from *A. thaliana*, *O. sativa,* and *M. truncatula* annotated in the most general molecular function GO categories revealed no significant conspicuities. Differences in the category “binding” and “catalytic activity” are nothing out of the ordinary, because *A. thaliana* has a very small genome size (~125Mbp) compared with *M. truncatula* (~454 to 526 Mbp) and *O. sativa* (~430 Mbp). One of the reasons for that difference could be whole genome duplication events [Goff et al. 2002]. But the absolute number of genes annotated with “signal transducer activity” is increased in Medicago in comparison with Arabidopsis. Most of the genes annotated in this category could be identified as disease resistance genes or Mal/TIRAP like genes, which are involved in the innate and adaptive immunity. This could indicate that in the Medicago genome more disease resistance related genes are present as in Arabidopsis, but because at this time (October 2006) the assembly of the Medicago DNA sequences is not finished, some of the genes could be redundant in the initial genome set.

In the Sorghum genome differences between Sorghum and its closest relative *O. sativa* could be discovered. However, because GO terms from *O. sativa* were not predicted by SIFTER, these differences might be caused by the difference in method. This could be corrected by running the pipeline with *O. sativa* as input and comparing the results to Sorghum.

In the case of the tomato genome project most GO categories are present in the batch11 set and therefore the distribution of the percentage of genes annotated to the general GO categories is quite representative for the whole tomato genome. However many genes in the GO categories “Cell killing”, “Growth”, “Pigmentation” and “Rhythmic process” are missing and the absolute number of genes in the categories is not comparable to other plants.

Significant differences were found by comparing the GO term annotations between animals and plant genomes. One difference is the number of genes annotated as signal transducer activity, which is increased in animals. But this must not necessarily mean that there is more signal transducer activity present in animals than in plants, because genes in this category are well explored in animals (like neurotransmitter activity and receptor activity in brain), but not so well known in plants. A further difference includes the category “nutrient reservoir activity”. I assume that all animal genes annotated with this category are wrongly annotated, because all rat and human genes annotated with that category are kinases and have a very short InterPro domain (InterPro ID:IPR000480) annotated, which has the GO term GO:0045735 (“nutrient reservoir activity”) assigned. This GO term is not experimentally verified in any animal protein. Furthermore *Mus musculus* proteins annotated with that GO term are phospholipases and the annotation was removed from the genes in future annotation files.

Non-plant specific genes

By looking for genes in the Medicago genome, which have homologous genes in any organism except plants, we could identify one topoisomerase, which seems to be a horizontal gene transfer, and three transferrin-like genes, which were not known in higher plants yet and are likely to belong to the same transferrin subfamily as insect and algal transferrin-like proteins. Further plant transferrins could be found in *Citrus clementina*, *Picea glauca, Picea sitchensis, Pinus taeda, Amborella trichopoda, Adiantum capillus-veneris* and *Pseudotsuga menziesi.* Because of their similarity to insect and algal transferrin-like proteins, I propose that these proteins in higher plants could play a role in the innate immune response against bacteria and fungi. Iron deprivation, as a result of iron binding proteins like transferrins, prevents the formation of a bacterial biofilm and makes bacteria nonresistant to innate immune defense or antibiotics [Ong et al. 2006]. In some insects it is shown that transferrin-like genes are up-regulated during infection [Valles and Pereira 2005] [Thompson et al. 2003] and in Drosophila they have been shown to be primarily dependent on the Toll-pathway, which is an important iron-withholding strategy [Boutros et al. 2002]. To prove this assumption further investigation is needed.

The phylogenetic tree of plant, insects, algae, cyanobacteria and sea urchins reflects the evolutionary history from primitive old organisms (e.g. algae, cyanobacteria or ferns) to higher evolved, younger organisms like Angiosperms and insects. Therefore I suppose that transferrins are a very old gene family and the first transferrin protein came from a primitive, ancient organism living in the sea. Lambert et al. assumed that the ancient transferrin may have only one transferrin domain [Lambert et al. 2005]. All transferrin-like genes in higher plants have only one transferrin domain and so this domain could be a direct descendant of the ancient transferrin gene. The duplication of that domain seems to be species family specific, because algae have three domains and insects and mammals have two transferrin domains.

In all other publicly available plant genomes (including mitochondria and chloroplast) no hints to transferrins or transferrin pseudo genes could be found. This raises the question why transferrin-like proteins seem to be only included in some organisms. One explanation for that could be a selective advantage, but till now it is unclear what kind of advantage this could be. Also in this case further experiments could give answers to that question.

In Sorghum I identified 14 non-plant-specific genes, which had no best hit in any plant or had no good hits in the genome of the plant model organism *A. thaliana*. Three of them are very interesting, because they are putative horizontal gene transfers from bacteria or come from mitochondrium or chloroplast. Also a calcium binding protein, a putative F-Box protein and an unknown protein could be identified, which are also found in rice, *Vitis vinifera*, Populus and *Picea sitchensis.* The functions of these genes are also unknown in other plants and it would be interesting to find out how they have evolved. Another three proteins did not have hits in any plants, but show low similarity to human genes. For all of them no protein domain could be detected and they seem to be unknown proteins in plants. In this case it has to be proven if these genes are expressed in Sorghum.

However, four of the 14 identified non-plant-specific proteins in Sorghum seem to be annotation errors, which means that splice sites are wrong and/or genes are not complete. The reason for that could be that gene prediction tools used for the gene calling in Sorghum were not trained with Sorghum genes and there was no manual verification of the annotated genes afterwards. If annotation errors are detected in the beginning these genes should be excluded from the function prediction, because they lead to errors.

VI. An accurate phylogenomic tool for automatic function prediction

1) Introduction

Besides the introduction of automatic pipelines for the functional predictions of genes, an accurate functional prediction tool is needed. Accurate means on the the one hand that the sensitivity should be as high as possible, which denotes that the set of annotated functions is most comprehensive, while having on the other hand a very high specificity, which implies that most of the predicted functions are true.

I have shown that SIFTER performs very well in the prediction of molecular function Gene Ontology terms for *M. truncatula* and *S. bicolor* genes (see chapter V). With the introduction of a pipeline to build a most comprehensive phylogenetic tree, problems of SIFTER regarding the tree topology could be solved and therefore the number of false predictions could be decreased. However, if only few genes in the phylogenetic tree have molecular function GO terms assigned, SIFTER is not able to distinguish between functionally related genes and genes with different functions (see chapter V). Furthermore SIFTER only considers molecular function GO terms at the lowest level of the GO graph as candidate functions. Sister nodes of these terms are not considered. However, in comparison to animal genes, which have in many cases low-level GO terms annotated, GO terms annotated to plant genes are often more general [Kourmpetis et al. 2007]. This can result in wrong predictions if the plant GO term is a parent of the animal GO term, but the plant gene has actually not the same function as the animal gene. Furthermore the set of annotated functions is for many genes incomplete, which complicates an comprehensive prediction of all functions [Kourmpetis et al. 2007]. Another problem is that SIFTER is just able to predict molecular function GO terms. But because each ontology has its own strengths and weaknesses it would make sense to predict terms from more than one function ontology. Additionally it is important to consider more than one candidate function for one gene [Kourmpetis et al. 2007].

In the following chapter I will introduce an extended, more accurate version of SIFTER (SIFTER-X), which uses additional functional attributes like domain information, interaction partners and different ontology terms annotated to genes in the phylogenetic tree to calculate a functional mutation rate. This functional mutation rate is used to either slow down the SIFTER mutation (see chapter III1b) in case of same attributes and speed up the SIFTER mutation in case of different attributes. In addition to GO molecular function SIFTER-X is also able to predict GO biological process terms, EC numbers [Webb et al. 1992], MapMan bins [Thimm et al. 2004] and KEGG ontology terms [Kanehisa et al. 2008]. To compare SIFTER and SIFTER-X and to calculate the accuracy of SIFTER-X both tools were tested on the photolyase/blue-light photoreceptor family and on a curated test set of 232 *A. thaliana* genes.

Photolyases are involved in UV-damaged DNA repair and are present in many species. Blue-light photoreceptors, also known as cryptochromes, regulate growth and development in plants and the circadian clock in animals. They are related to photolyases but have no photoreactivation activity and they are not involved in DNA repair [Malhotra et al. 1995] [Sancar 2003] [Hsu et al. 1996]. It is shown for the Cryptochrome1 in *A. thaliana* (CRY1) that the photoreceptor activity requires a light-induced homodimerisation of the N-terminal CNT1 domains of CRY1 [Sang et al. 2005]. Of all blue-light receptor genes in plants only the genes in Arabidopsis are functionally well characterized with ontology terms. Because of that and because these genes show a high sequence similarity to photolyase genes, they are often wrongly annotated. Other plant cryptochromes were discovered, among others, in *Brassica napus* (UniProt:Q1JU52_BRANA) [Chatterjee et al. 2006], *Nicotiana sylvestris* (UniProt:Q309E8_NICSY) [Yendrek and Metzger 2005], *Solanum lycopersicum* (Q9XHD8_SOLLC & Q93VS0_SOLLC) [Ninu et al. 1999] [Perrotta et al. 2000] and *Pisum sativum* (Q6YBV9_PEA, Q6EAN1_PEA) [Platten et al. 2005]. All plant cryptochromes show a high sequence conservation to the cryptochrome genes from *A. thaliana* (>70% identity and >80% positives) and share the same domain composition.

A. thaliana genes were chosen as the second test set, because *A. thaliana* is the best studied organism in plants and therefore most genetic and functional data is available for that organism. Currently most plant genes get functional annotations by comparison with *A. thaliana* genes. At the moment (status 11/2008) 91934 of the 112153 GO terms annotated to Arabidopsis genes are experimentally verified or curated [Gene Ontology 2008].

2) Materials & Methods

a) Collecting additional functional attributes available for genes in the phylogenetic tree

Additional functional attributes have been collected from different sources by using web services provided by different institutes. To integrate them into the so-called PLI file, an XML file which is used as input for SIFTER and includes all genes in the phylogenetic tree together with their functional annotations, the PLI file was extended by the XML elements described in table 13 and Java classes have been written to parse it within the SIFTER-X framework.

XML Element	*Children elements*	*Former XML element*	*Description*
GONumberMF	Term, Evidence	GONumber	Annotated molecular function GO term. Includes GO id and GO evidence code
GONumberBP	Term, Evidence	-	Annotated biological process GO term. Includes GO id and GO evidence code
KONumber	Term, Evidence	-	Annotated KO term. Includes KO term and KO evidence code
ECNumber	Term, Evidence	-	Annotated EC number. Includes EC number and evidence code
MapManBin	Term, Evidence	-	Annotated MapMan bin. Includes MapMan bin code and evidence code
Domain	Accession	-	Accession number for the protein domain included in the gene (e.g. InterPro ID)
Interaction	InteractionPartner	-	Physical interaction partner of the gene.

Table 13: New XML elements of the new PLI file used as input for SIFTER-X.

To enable a fast and easy PLI file generation, a program has been implemented which reads database accessions from the phylogenetic tree, runs different web services to get the additional information and includes this information in the PLI file. To get additional database accessions for all genes and thereby enlarge the number of web services which can be used, the mapping service PICR from the European Bioinformatics Institute [Côté et al. 2007] is called in the beginning.

Physical interaction partners and interacting molecules for each gene are retrieved from the IntAct web service [Hermjakob et al. 2004] and the BIND web service [Bader et al. 2003], whereas the DBFetch web service [Labarga et al. 2007] is run to get the UniProt entry [UniProt 2007] for each gene, from which InterPro identifiers [Hunter et al. 2008] and EC numbers are parsed. The KEGG web service bconv [Kanehisa et al. 2008] is used to get KEGG database identifiers for each gene which are used to call the get_ko_by_gene web service [Kanehisa et al. 2008] to get KO terms. MapMan bins are fetched via the BioMOBY web services getBinCodeByUniProtKB_id and getBinCodeByAGI provided by the MapMan consortium [Thimm et al. 2004]. GO terms for molecular function and biological process are collected from the BioMOBY getGOTermByDatabaseID web service provided at the Max-Planck Institute for Plant Breeding Research. This web service provides GO terms for genes from annotation files provided at the Gene Ontology website[14].

b) Extension of the SIFTER algorithm (SIFTER-X)

For SIFTER-X only the propagation step without the maximum likelihood approach has been extended [Engelhardt et al. 2005], because running SIFTER with the maximum likelihood setting is not suitable for genome projects due to the long running time. The SIFTER algorithm was extended in three ways. As the first step we included the functional mutation rate (FMR) to change the mutation rate within the SIFTER framework, which in the former SIFTER version relies only on the branch length between nodes in the tree and the evolutionary event that occurred at this node (if it is a duplication or a speciation node) (see chapter III1b). The FMR is calculated separately for each function set (MapMan bin, EC number, GO molecular function, GO biological process, KO term, protein domains, physical interaction partners) and the overall functional mutation rate at each intermediate node in the tree is the average of all calculated FMRs.

To calculate the FMR for one function set, each node in the tree is expressed as a vector in which each position indicates the frequency of a functional attribute in the descendant tree of this node or in case of a leaf the number of a functional attribute available for a certain gene. Each vector is normalized afterwards to length 1 by applying equation 2 to all vector elements x_i.

$$\tilde{x}=\frac{x_i}{\sqrt{\sum_{j=1}^{n} x_j}}$$

Equation 2: Function set vectors x are normalized to length 1.

The FMR at node M for a certain function set is calculated in case of ontology terms by using the euclidean distance between the vectors of the children nodes X and Y of M (see equation 3). To take care of parent-child relationships between ontology terms, vector elements which are 0 are replaced

[14] *http://www.geneontology.org/GO.current.annotations.shtml*

by the frequency of the related term if there is a parent-child relationship between the corresponding ontology terms.

$$Euclidean\ distance(\langle x\rangle,\langle y\rangle) = \sqrt{\sum (x_i - y_i)^2}$$

Equation 3: Euclidean distance between vector x and vector y.

To compute the functional mutation rate for the protein domains the maximum distance is used (see equation 4) instead of the euclidean distance, because missing domains are often linked to change of function [Kriventseva et al. 2003] and the maximum distance weights differences higher than similarities.

$$Maximum\ distance(\langle x\rangle,\langle y\rangle) = max(|x_i - y_i|)$$

Equation 4: Maximum distance between vector x and vector y.

In case of the functional mutation rate of interaction partners it is just considered if at least one of the interaction partners assigned to the children nodes is equal for both children nodes or all interaction partners are different. This approach is based on the assumption that some interaction partners are not discovered yet and the error rate is high [Grigoriev 2003]. Because each database, which includes physical interaction data, returns different database accessions for interacting proteins, all interactions were first classified in four classes according to the available database accessions:

- Small molecule names
- GI number
- Arabidopsis locus code identifier (AGI)
- UniProt ID

To enable a fast computation of the functional mutation rate within the SIFTER-X framework, database accessions are not mapped to each other and each set of accessions is considered separately as one function set. If at least one interaction partner is equal for both children nodes, the functional mutation rate is 2, because I assume that these nodes have same or overlapping functions. If all interaction partners are different, the functional mutation rate is 0. Only interaction partners of the same organism are considered.

At the end, the FMR calculated by the euclidean or maximum distance is further normalized to a value in range 0 to 2 and the FMR for all function sets (average of all FMRs) is multiplied with the former mutation rate based on branch length and evolutionary event (duplication/speciation) (see chapter III1b). A value for the FMR in the range 0 to 2 is chosen, so that in case of completely equal functional attributes the FMR is 0 and the mutation rate between nodes becomes 0, which means that functions can be transferred between nodes and both branch length and the evolutionary event is not taken into account anymore. In the other case, if all functional attributes for each function set are different, the FMR is 2, which results in doubling the mutation rate and so it becomes more unlikely that the function is transferred between nodes.

As the second step classes to parse the KEGG ontology, the EC number hierarchy, the whole Gene Ontology graph and the MapMan bin hierarchy have been implemented and parts of SIFTER have been rewritten to be more general, to be able to predict other ontology terms and to make use of the relationships between terms for calculating the FMR. In SIFTER-X the user has the possibility to decide which ontology should be taken for the prediction process.

As the last step the pruning of the GO graph was changed in the way that nodes are not removed if one of their sister nodes is annotated to a protein inside the phylogenetic tree. In addition to that all annotated ontology terms are considered as candidate functions for all genes to deal with incomplete ontologies and non-specific ontology terms annotated to genes. To reduce the complexity of the SIFTER-X output only the most probable ontology term is given as output for the corresponding gene if two ontology terms are in a parent-child relationship.

c) Building the first test set "The Blue-Light Photoreceptor/Photolyase family"

The *A. thaliana* blue-light receptor Cryptochrome1 (RefSeq ID: NP_567341.1) was used as input for the iterative Blast search described in chapter V2a to search for in-paralogous and orthologous genes. Because the former database does not include all organisms, a WU-Blast search (version: 2.0MP-WashU [04-May-2006]) [Gish 1996-2004] was run against the UniProt database (release 14.0) [UniProt 2007] and the ten best hits were extracted and integrated in the set of family members to increase the number of plant blue light photoreceptor genes in the family. After removing redundant genes, a multiple sequence alignment of all candidate family members (see appendix) was built using MAFFT (Version: 6.24) [Katoh et al. 2005] and a bootstrapped parsimony tree was built using seqboot, protpars and consense from the PHYLIP package (Version: 3.65) [Felsenstein 1993]. To find out if the tree topology has any influence on the overall SIFTER and SIFTER-X result, the tree was re-rooted manually afterwards to connect the true root of the tree with the middle of the branch between photolyase family and cryptochrome family.

I assume that all *Medicago truncatula, Oryza sativa, Vitis vinifera and Brassica campestris* genes in the tree, which share the same subtree with the known plant cryptochromes in *A. thaliana, Brassica napa, Nicotiana sylvestris, Solanum lycopersicum and Pisum sativum* are also cryptochromes, because they have a high sequence similarity to the known plant cryptochromes (see figure 15) and share the same protein domains.

d) Building a curated data set of A. thaliana genes

To test SIFTER-X on a well curated gene dataset, we decided on a test set of 232 randomly chosen *A. thaliana* proteins which have at least one experimentally verified or curated molecular function GO term assigned (see appendix) and for which the phylogenomic pipeline with SIFTER (see chapter V2b) was able to predict at least one molecular function GO term. Because on the one hand the predicted GO terms were more specific than the annotated terms and on the other hand the predicted terms do not even occur in the annotation, all predicted functions by SIFTER and SIFTER-X were manually checked afterwards by using sequence comparison and published literature. For this the same criteria were used as for the Medicago dataset. (described in chapter V2e).

e) Applying SIFTER and SIFTER-X on the test datasets

To apply SIFTER and SIFTER-X on the test datasets the pipeline described in chapter V2b was modified by running SIFTER and SIFTER-X with the following arguments:

- SIFTER arguments to generate family, scale and alpha files:

  --familyfile <FILE> --scale <FILE> --alpha <FILE> --with-ic –with-iep –with-igi –with-ipi –with-iss –with-rca –with-tas –with-nas –generate –reconciled <TREE FILE> --ontology <function.ontology FILE> --protein <PLI FILE> FAMILYNAME

- SIFTER arguments to predict GO terms (using generated family, scale and alpha files):

  --familyfile <FAMILY FILE> --scale <SCALE FILE> --alpha <ALPHA FILE> --with-ic –with-iep –with-igi –with-ipi –with-iss –with-rca –with-tas –with-nas –generate –reconciled <TREE FILE> --ontology <function.ontology FILE> --output <SIFTER OUTPUT FILE> --protein <PLI FILE> --truncation 2 FAMILYNAME

- SIFTER-X arguments

  --use-curated –reconciled <TREE FILE> --ontology <ONTOLOGY TO USE> --output <SIFTER OUTPUT FILE> --protein <PLI FILE> --truncation 2 FAMILYNAME

For the Arabidopsis test set all ontology term annotations to the query proteins were removed in the beginning. While for the blue-light photoreceptor/photolyase family the predicted GO terms for all genes in the tree were taking into account, for the curated Arabidopsis gene set only predicted GO terms of the query Arabidopsis gene were used. SIFTER was modified so that all predicted GO terms are printed into the output file.

f) Evaluation of the SIFTER and SIFTER-X results

$$Sensitivity = \frac{TP}{TP+FN}$$

Equation 5: The Sensitivity is the ratio of True Positives (TP) over the sum of TPs and False Negatives (FN).

$$Specificity = \frac{TN}{TN+FP}$$

Equation 6: The Specificity is the ratio of True Negatives (TN) over the sum of TNs and False Positives (FP).

To calculate the sensitivity (see equation 5) and the specificity (see equation 6) of SIFTER and SIFTER-X for each posterior probability cutoff a Java program was written which iterates through the list of all predicted ontology terms and their corresponding posterior probabilities and checks if the predicted ontology term is present in the set of true ontology terms available for each gene (see chapter VI2d) or not. The following criteria were used to separate True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) for any predicted function F by applying a certain cutoff:

TP: Gene has function F (or any child term of F) and posterior probability >= cutoff

TN: Gene has not function F and posterior probability < cutoff

FP: Gene has not function F and posterior probability >= cutoff

FN: Gene has function F and posterior probability < cutoff or true function is not predicted by the analysis program

g) Evaluation of the Blast results

Blast does not predict functions, but often the functional annotation of the best Blast hit or of all hits with an e-value higher than a certain cutoff are transferred to the query gene. To calculate the sensitivity (see equation 5) and the specificity (see equation 6) for both approaches a NCBI-Blast (Version: 2.2.13) was run against the manually built RefSeq database described in chapter V2a using all genes in the curated Arabidopsis test set as query (see chapter VI2d). A Java program parses the Blast result, gets for all hits the molecular function GO term and the minimum e-value and counts the number of True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) GO terms by applying a certain e-value cutoff. The following criteria were used to count TPs, TNs, FPs and FNs:

TP: Gene has function F (or any child term of F) and e-value <= e-value cutoff

TN: Gene has not function F and e-value > e-value cutoff

FP: Gene has not function F and e-value <= e-value cutoff

FN: Gene has function F and e-value > e-value cutoff or true function is not predicted by the analysis program

The sensitivity and specificity was calculated for two scenarios. In the first case molecular function GO terms of all hits better than a given e-value cutoff are considered. In the second case only the best hit is taking into account, which has a molecular function GO annotation. Self hits were ignored.

3) Results

a) Application I: The Blue-Light Photoreceptor/Photolyase family

We tested SIFTER and SIFTER-X using the same phylogenetic tree as input on the blue-light photoreceptor/photolyase family [Kanai et al. 1997]. Blue-light photoreceptor genes are often wrongly annotated by other tools, because they share four of five protein domains with photolyase genes and three of them (IPR002081, IPR006050,IPR005101) have GO term GO:0003913 (photolyase activity) and GO term GO:0006281 (DNA repair) assigned from InterPro2GO. However known blue-light photoreceptor genes have no photolyase activity and are not involved in DNA repair [Malhotra et al. 1995] [Sancar 2003] [Hsu et al. 1996].

Prediction of molecular function GO terms and comparison with SIFTER

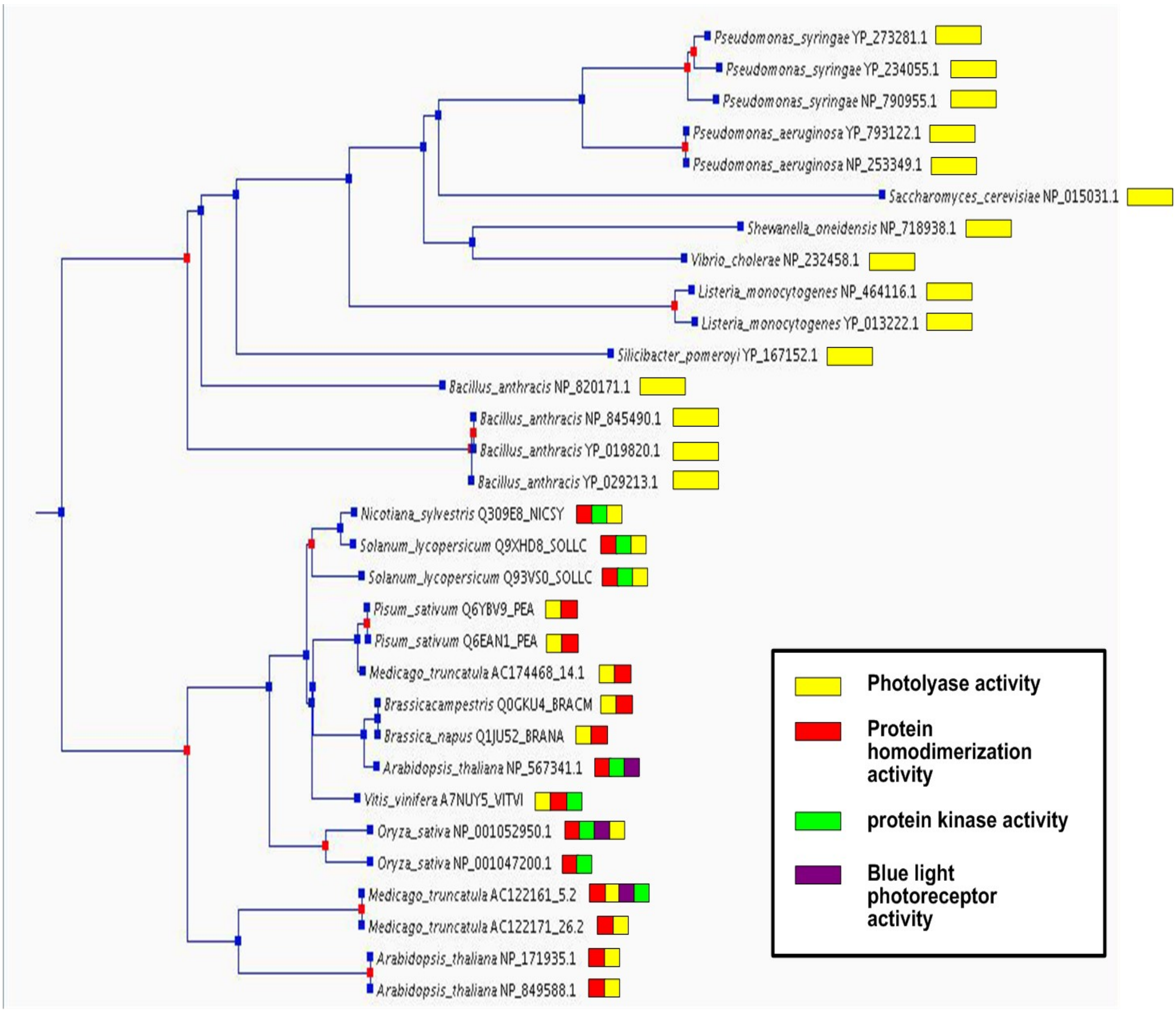

Pseudomonas_syringae YP_273281.1
Pseudomonas_syringae YP_234055.1
Pseudomonas_syringae NP_790955.1
Pseudomonas_aeruginosa YP_793122.1
Pseudomonas_aeruginosa NP_253349.1
Saccharomyces_cerevisiae NP_015031.1
Shewanella_oneidensis NP_718938.1
Vibrio_cholerae NP_232458.1
Listeria_monocytogenes NP_464116.1
Listeria_monocytogenes YP_013222.1
Silicibacter_pomeroyi YP_167152.1
Bacillus_anthracis NP_820171.1
Bacillus_anthracis NP_845490.1
Bacillus_anthracis YP_019820.1
Bacillus_anthracis YP_029213.1
Nicotiana_sylvestris Q309E8_NICSY
Solanum_lycopersicum Q9XHD8_SOLLC
Solanum_lycopersicum Q93VS0_SOLLC
Pisum_sativum Q6YBV9_PEA
Pisum_sativum Q6EAN1_PEA
Medicago_truncatula AC174468_14.1
Brassicacampestris Q0GKU4_BRACM
Brassica_napus Q1JU52_BRANA
Arabidopsis_thaliana NP_567341.1
Vitis_vinifera A7NUY5_VITVI
Oryza_sativa NP_001052950.1
Oryza_sativa NP_001047200.1
Medicago_truncatula AC122161_5.2
Medicago_truncatula AC122171_26.2
Arabidopsis_thaliana NP_171935.1
Arabidopsis_thaliana NP_849588.1
Photolyase activity
Protein homodimerization activity
protein kinase activity
Blue light photoreceptor activity

Figure 15: SIFTER molecular function Gene Ontology annotation for the photolyase/blue-light photoreceptor family. Yellow, red, green and purple rectangles indicate which GO terms have been predicted with the best (first position from left to right), second best, third best and fourth best posterior probability. A cutoff of >=0.1 was used. This figure was taken from Jöcker et al. 2009.

SIFTER was not able to distinguish between the different families and assigned in 6 of 16 cases (37,5%) GO:0003904 (deoxyribodipyrimidine photo-lyase activity) with the best posterior probability (see green colored predictions of SIFTER in table 15) and in four cases with the second best posterior probabilities (see light blue colored predictions of SIFTER in table 15) to the blue-light photoreceptor genes (see yellow colored boxes in figure 15).

Only for three proteins (see light red colored predictions of SIFTER in table 15 and figure 15) SIFTER was able to predict the right functions (GO:0042803 (protein homodimerization activity), GO:0009882 (blue light photoreceptor activity) and GO:0004672 (protein kinase activity)) with the best posterior probability. However only in one case (NP_567341.1 (CRY1 from *Arabidopsis thaliana*)) there was a significant difference of more than 0.1 between one of the true GO terms and the wrong GO term GO:0003904 due to the fact that GO:0004672, GO:0009882 and GO:0042803 were already annotated to this protein with the evidence codes IDA (Inferred from Direct Assay), IMP (Inferred from Mutant Phenotype) and IPI (Inferred from Physical Interaction) respectively and so these GO terms get for this node a very high initial probability. In comparison to that all photolyase proteins received the right GO term GO:0003904 with a probability greater than 97%.

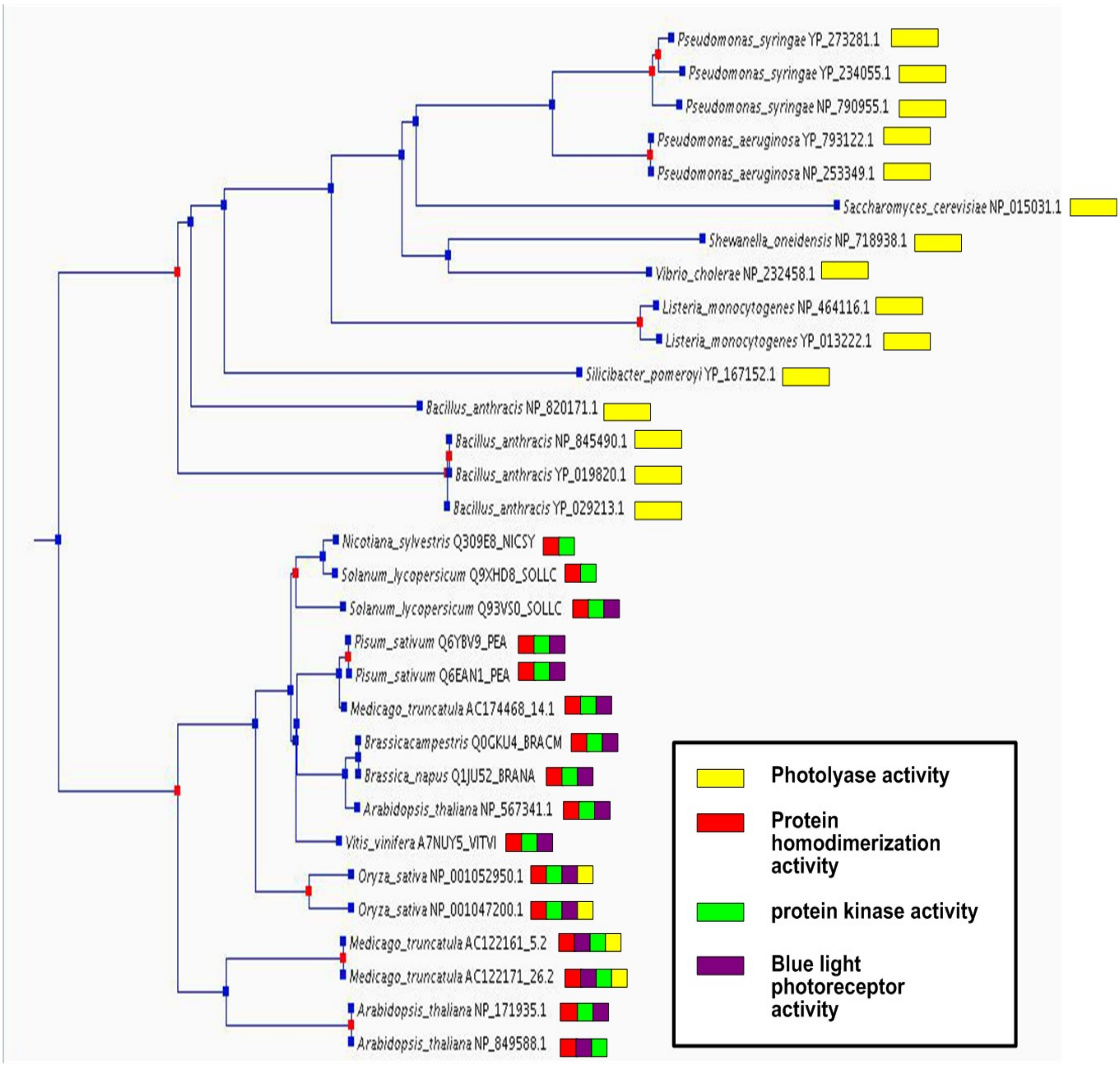

Pseudomonas_syringae YP_273281.1
Pseudomonas_syringae YP_234055.1
Pseudomonas_syringae NP_790955.1
Pseudomonas_aeruginosa YP_793122.1
Pseudomonas_aeruginosa NP_253349.1
Saccharomyces_cerevisiae NP_015031.1
Shewanella_oneidensis NP_718938.1
Vibrio_cholerae NP_232458.1
Listeria_monocytogenes NP_464116.1
Listeria_monocytogenes YP_013222.1
Silicibacter_pomeroyi YP_167152.1
Bacillus_anthracis NP_820171.1
Bacillus_anthracis NP_845490.1
Bacillus_anthracis YP_019820.1
Bacillus_anthracis YP_029213.1
Nicotiana_sylvestris Q309E8_NICSY
Solanum_lycopersicum Q9XHD8_SOLLC
Solanum_lycopersicum Q93VS0_SOLLC
Pisum_sativum Q6YBV9_PEA
Pisum_sativum Q6EAN1_PEA
Medicago_truncatula AC174468_14.1
Brassicacampestris Q0GKU4_BRACM
Brassica_napus Q1JU52_BRANA
Arabidopsis_thaliana NP_567341.1
Vitis_vinifera A7NUY5_VITVI
Oryza_sativa NP_001052950.1
Oryza_sativa NP_001047200.1
Medicago_truncatula AC122161_5.2
Medicago_truncatula AC122171_26.2
Arabidopsis_thaliana NP_171935.1
Arabidopsis_thaliana NP_849588.1
Photolyase activity
Protein homodimerization activity
protein kinase activity
Blue light photoreceptor activity

Figure 16: SIFTER-X molecular function Gene Ontology annotation for the photolyase/blue-light photoreceptor family. Yellow, red, green and purple rectangles indicate which GO terms have been predicted with the best (first position from left to right), second best, third best and forth best posterior probability. A cutoff of >=0.1 was applied. This figure was taken from Jöcker et al. 2009.

However, there is one protein domain (IPR014134) detected by InterProScan which is not found in the photolyase family. Furthermore in the blue-light photoreceptor family all Arabidopsis genes (NP_567341.1 (Cryptochrome 1) & NP_171935.1/NP_849588.1 (Cryptochrome 2)) have the same MapMan bin (30.11 (signalling.light)) assigned and share the same interaction partners (AT2G32950, AT2G18790). In the photolyase family all members have a KO term assignment (K01669) and some also have the EC number 4.1.99.3 (deoxyribodipyrimidine photo-lyase) assigned. Because of this additional information SIFTER-X was able to distinguish between the different families and assigned term GO:0003904 with a high posterior probability (0.99) to all photolyase genes and with the lowest posterior probability to all blue-light photoreceptor genes (see table 15 and figure 16). In 10 out of 14 cases of the blue-light photoreceptor genes (see yellow colored results in table 15 and figure 16) the predicted posterior probability for all true GO terms was higher than 0.1 and for the wrong GO terms below 0.1. However, SIFTER-X predicted GO:0009882 (blue light photoreceptor activity) for 10 proteins with a probability smaller than 0.2. The reason for that was that GO:0009882 was annotated to just two proteins (NP_171935.1 and NP_567341.1). For protein NP_171935.1 GO:0009882 was annotated with the evidence code ISS (Inferred from Sequence or Structural Similarity), which means that the function was reviewed, but not experimentally verified and therefore was initialized with a very low probability of 0.4 to be true. NP_567341.1 had GO:0009882 annotated with IMP (Inferred from Mutant Phenotype), which got an initial probability of 0.8.

GO term	***GO term name***	***Cryptochrome1***	***Cryptochrome2***
0009638	Phototrophism	X	√
0009414	Response to water deprivation	√	√
0010118	Stomatal movement	√	√
0009637	Response to blue light	√	√
0009909	Regulation of flower development	X	√
0009911	Positive regulation of flower development	X	√
0006338	Chromatin remodeling	X	√
0009785	Blue light signaling pathway	√	X
0009640	Photomorphogenesis	√	X
0046777	Protein amino acid autophosphorylation	√	X
0006118	Transport	√	X
007623	Circadian rhythm	√	X
0046283	Antocyanin metabolic process	√	X

Table 14: Annotated biological process GO terms for A. thaliana Cryptochrome1 and Cryptochrome2. A cross indicates that this GO term was not annotated to that protein. This table was taken from Jöcker et al. 2009.

The SIFTER-X result for the blue-light photoreceptor proteins can be slightly improved (see table 15) by excluding the biological process GO terms for the functional mutation rate calculation, because *A. thaliana* Cryptochrome1 and Cryptochrome2 share only 3 of 13 biological process GO terms (see table 14). After excluding the biological process GO terms for the functional mutation rate calculation the posterior probability predicted for all true GO terms was higher than 0.1 and for the wrong GO term GO:0003904 was in 12 cases smaller than 0.1.

Protein name (database)	***Predicted GO term by SIFTER***	***Predicted posterior probability by SIFTER***	***True?***	***Predicted GO term by SIFTER-X***	***Predicted posterior probability by SIFTER-X***	***Predicted posterior probability by SIFTER-X (excluding biological process GO terms)***	***True?***
Q309E8_NICSY (UniProt)	0003904 0004672 0009882 0042803	0.19 0.30 0.08 0.88	No Yes Yes Yes	0003904 0004672 0009882 0042803	0.0004 0.40 0.08 0.95	0.05 0.36 0.11 0.93	No Yes Yes Yes
Q9XHD8_SOLLC (UniProt)	0003904 0004672 0009882 0042803	0.18 0.28 0.07 0.80	No Yes Yes Yes	0003904 0004672 0009882 0042803	0.0004 0.40 0.08 0.95	0.05 0.36 0.11 0.93	No Yes Yes Yes
Q93VS0_SOLLC (UniProt)	0003904 0004672 0009882 0042803	0.18 0.28 0.07 0.80	No Yes Yes Yes	0003904 0004672 0009882 0042803	0.0007 0.41 0.11 0.91	0.08 0.39 0.16 0.89	No Yes Yes Yes
Q6YBV9_PEA (UniProt)	0003904 0004672 0009882 0042803	0.96 0.08 0.07 0.32	No Yes Yes Yes	0003904 0004672 0009882 0042803	0.07 0.15 0.14 0.99	0.03 0.22 0.11 0.99	No Yes Yes Yes
Q6EAN1_PEA (UniProt)	0003904 0004672 0009882 0042803	0.99 0.02 0.01 0.26	No Yes Yes Yes	0003904 0004672 0009882 0042803	0.07 0.15 0.14 0.99	0.03 0.22 0.11 0.99	No Yes Yes Yes
AC174468_14.1 (IMGAG 1.0)	0003904 0004672 0009882 0042803	0.95 0.07 0.06 0.30	No Yes Yes Yes	0003904 0004672 0009882 0042803	0.06 0.15 0.14 0.98	0.009 0.22 0.11 0.98	No Yes Yes Yes
Q0GKU4_BRACM (UniProt)	0003904 0004672 0009882 0042803	0.90 0.07 0.07 0.38	No Yes Yes Yes	0003904 0004672 0009882 0042803	0.03 0.20 0.14 0.98	0.03 0.17 0.13 0.99	No Yes Yes Yes
Q1JU52_BRANA (UniProt)	0003904 0004672 0009882 0042803	0.90 0.07 0.06 0.38	No Yes Yes Yes	0003904 0004672 0009882 0042803	0.03 0.20 0.14 0.98	0.03 0.17 0.13 0.99	No Yes Yes Yes
NP_567341.1 (RefSeq protein)	0003904 0004672 0009882 0042803	0.08 0.33 0.19 0.56	No Yes Yes Yes	0003904 0004672 0009882 0042803	0.0004 0.45 0.39 0.99	0.0004 0.40 0.34 0.99	No Yes Yes Yes
A7NUYS_VITVI (UniProt)	0003904 0004672	0.52 0.10	No Yes	0003904 0004672	0.001 0.22	0.002 0.38	No Yes

Protein name (database)	***Predicted GO term by SIFTER***	***Predicted posterior probability by SIFTER***	***True?***	***Predicted GO term by SIFTER-X***	***Predicted posterior probability by SIFTER-X***	***Predicted posterior probability by SIFTER-X (excluding biological process GO terms)***	***True?***
	0009882 0042803	0.04 0.46	Yes Yes	0009882 0042803	0.12 0.93	0.14 0.89	Yes Yes
NP_001052950.1 (RefSeq protein)	0003904 0004672 0009882 0042803	0.10 0.48 0.12 0.77	No Yes Yes Yes	0003904 0004672 0009882 0042803	0.17 0.45 0.25 0.78	0.12 0.44 0.20 0.85	No Yes Yes Yes
NP_001047200.1 (RefSeq protein)	0003904 0004672 0009882 0042803	0.06 0.45 0.07 0.77	No Yes Yes Yes	0003904 0004672 0009882 0042803	0.17 0.45 0.25 0.78	0.12 0.44 0.20 0.85	No Yes Yes Yes
AC122161_5.2 (IMGAG 1.0)	0003904 0004672 0009882 0042803	0.25 0.12 0.15 0.90	No Yes Yes Yes	0003904 0004672 0009882 0042803	0.12 0.25 0.26 0.84	0.11 0.42 0.21 0.74	No Yes Yes Yes
AC122171_26.1 (IMGAG 1.0)	0003904 0004672 0009882 0042803	0.19 0.06 0.09 0.91	No Yes Yes Yes	0003904 0004672 0009882 0042803	0.12 0.25 0.26 0.84	0.11 0.42 0.21 0.75	No Yes Yes Yes
NP_171935.1 (RefSeq protein)	0003904 0004672 0009882 0042803	0.24 0.04 0.05 0.25	No Yes Yes Yes	0003904 0004672 0009882 0042803	0.01 0.21 0.15 0.50	0.007 0.37 0.15 0.49	No Yes Yes Yes
NP_849588.1 (RefSeq protein)	0003904 0004672 0009882 0042803	0.42 0.08 0.08 0.50	No Yes Yes Yes	0003904 0004672 0009882 0042803	0.01 0.13 0.25 0.89	0.009 0.22 0.22 0.87	No Yes Yes Yes

Table 15: Molecular function GO term predictions made by SIFTER and SIFTER-X for proteins of the blue-light photoreceptor family. Posterior probabilities are rounded. The green color indicates that SIFTER has predicted the wrong GO term with the highest posterior probability. Light blue denotes that the wrong GO term was predicted by SIFTER with the second best probability and light red shows predictions made by SIFTER for which the wrong GO term got the lowest probability. Yellow colored boxes indicate predictions for which all true GO terms got a probability > 0.1 and all wrong annotated GO terms got a probabiltiy < 0.1. This table was taken from Jöcker et al. 2009.

Prediction of biological process GO terms

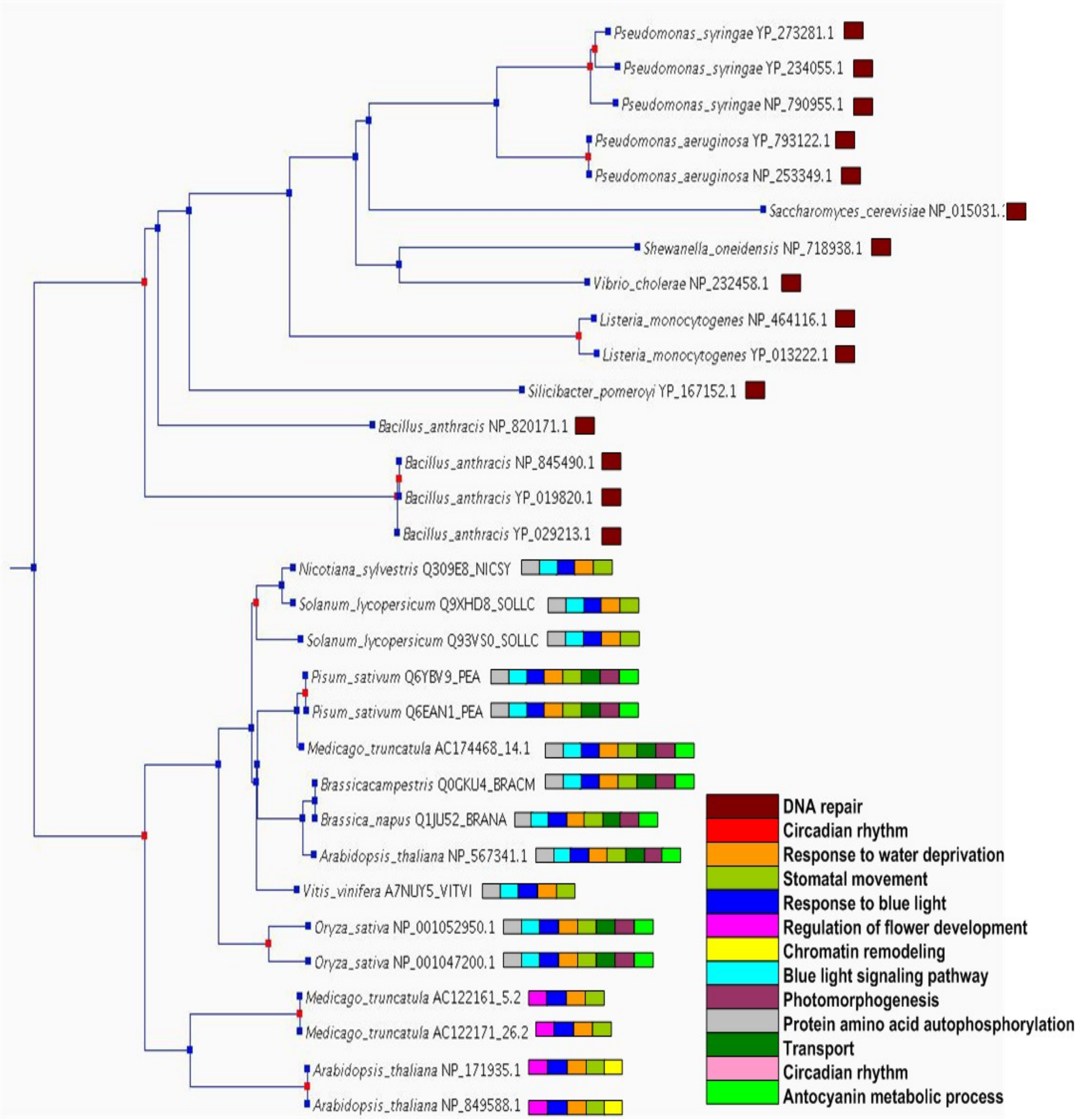

Pseudomonas_syringae YP_273281.1
Pseudomonas_syringae YP_234055.1
Pseudomonas_syringae NP_790955.1
Pseudomonas_aeruginosa YP_793122.1
Pseudomonas_aeruginosa NP_253349.1
Saccharomyces_cerevisiae NP_015031.1
Shewanella_oneidensis NP_718938.1
Vibrio_cholerae NP_232458.1
Listeria_monocytogenes NP_464116.1
Listeria_monocytogenes YP_013222.1
Silicibacter_pomeroyi YP_167152.1
Bacillus_anthracis NP_820171.1
Bacillus_anthracis NP_845490.1
Bacillus_anthracis YP_019820.1
Bacillus_anthracis YP_029213.1
Nicotiana_sylvestris Q309E8_NICSY
Solanum_lycopersicum Q9XHD8_SOLLC
Solanum_lycopersicum Q93VS0_SOLLC
Pisum_sativum Q6YBV9_PEA
Pisum_sativum Q6EAN1_PEA
Medicago_truncatula AC174468_14.1
Brassicacampestris Q0GKU4_BRACM
Brassica_napus Q1JU52_BRANA
Arabidopsis_thaliana NP_567341.1
Vitis_vinifera A7NUY5_VITVI
Oryza_sativa NP_001052950.1
Oryza_sativa NP_001047200.1
Medicago_truncatula AC122161_5.2
Medicago_truncatula AC122171_26.2
Arabidopsis_thaliana NP_171935.1
Arabidopsis_thaliana NP_849588.1
DNA repair
Circadian rhythm
Response to water deprivation
Stomatal movement
Response to blue light
Regulation of flower development
Chromatin remodeling
Blue light signaling pathway
Photomorphogenesis
Protein amino acid autophosphorylation
Transport
Circadian rhythm
Antocyanin metabolic process

Figure 17: SIFTER-X biological process Gene Ontology annotation for the photolyase/blue-light photoreceptor family. Colored rectangles indicate which GO terms have been predicted with the best (first position from left to right), second best, third best and so on best posterior probability. A cutoff of >=0.1 was applied. This figure was taken from Jöcker et al. 2009.

SIFTER-X was also tested on the prediction of biological process GO terms. As shown in figure 17 SIFTER-X was able to differentiate between three different subgroups in the tree. The first subgroup includes all photolyase genes (see upper part of figure 17) and all genes in this group were assigned the true GO term GO:0006281 (DNA repair) with a posterior probability higher than 0.98 assigned. Only one protein NP_015031.1 received this GO term with a lower probability of 0.73. The second subgroup incorporates Cryptochrome1 (NP_567341.1) of *A. thaliana* and putative orthologous genes from different organisms (Q309E8_NICSY, Q9XHD8_SOLLC, Q93VS0_SOLLC, Q6YBV_PEA, Q6EAN1_PEA, AC174468_14.1, Q0GKU4_BRACM, Q1JU52_BRANA, A7NUY5_VITVI, NP_001052950.1 and NP_001047200.1). When applying a posterior probability cutoff of 0.1, Q309E8_NICSY, Q9XHD8_SOLLC, Q93VS0_SOLLC and A7NUY5_VITVI were assigned the GO terms GO:0046777 (Protein amino acid autophosphorylation), GO:0009785 (Blue light signaling pathway), GO:0009637 (Response to blue light), GO:0009414 (Response to water deprivation) and GO:0010118 (Stomatal movement). However, the other genes in this group got three additional GO terms (GO:0006118 (Transport), GO:0009640 (Photomorphogenesis) and GO:0046283 (Antocyanin metabolic process)) annotated with a posterior probability higher than 0.1 (see table 17). The third subgroup in the tree consists of Cryptochome2 (NP_171935.1 & NP_849588.1) from *A. thaliana* and two putative orthologous genes in *M. truncatula* (AC122161_5.2 & AC122171_26.2). All genes in the third subgroup were assigned the GO terms GO:0009909 (Regulation of flower development), GO:0009637 (Response to blue light), GO:0009414 (Response to water deprivation) and GO:0010118 (Stomatal movement) with a posterior probability higher than 0.1. In addition to that Cryptochrome2 genes from *A. thaliana* were assigned GO term GO:0006338 (Chromatin remodeling) at a probability cutoff of 0.1.

Except for the Arabidopsis proteins (Cryptochrome I and II) which already had these GO terms assigned, the posterior probability for all annotated biological process GO terms to blue-light photoreceptor proteins is very low (<0.3) (see table 17), but their probability is significantly higher than for the wrong GO term GO:0006281 (DNA repair).

Subgroup	***Protein name***	***Predicted biological process GO terms***	***Posterior probability***
I	YP_273281.1	0006281 (DNA repair)	0.99
I	YP_234055.1	0006281 (DNA repair)	0.99
I	NP_790955.1	0006281 (DNA repair)	0.99
I	YP_793122.1	0006281 (DNA repair)	0.99
I	NP_253349.1	0006281 (DNA repair)	0.99
I	NP_015031.1	0006281 (DNA repair)	0.73
I	NP_718938.1	0006281 (DNA repair)	1.0
I	NP_232458.1	0006281 (DNA repair)	1.0
I	NP_464116.1	0006281 (DNA repair)	1.0
I	YP_013222.1	0006281 (DNA repair)	1.0
I	YP_167152.1	0006281 (DNA repair)	0.99
I	NP_820171.1	0006281 (DNA repair)	0.99
I	NP_845490.1	0006281 (DNA repair)	1.0
I	YP_019820.1	0006281 (DNA repair)	1.0
I	YP_029213.1	0006281 (DNA repair)	0.98

Table 16: Biological process GO term predictions by SIFTER-X for photolyase proteins (subgroup I) at a posterior probability cutoff of 0.1. This table was taken from Jöcker et al. 2009.

Subgroup	***Protein name***	***Predicted biological process GO terms***	***Posterior probability***
II	Q309E8_NICSY	0007623 (Circadian rhythm)	0.006
		0046777 (Protein amino acid autophosphorylation)	0.34
		0006281 (DNA repair)	0.0001
		0006118 (Transport)	0.08
		0009785 (Blue light signaling pathway)	0.34
		0009414 (Response to water deprivation)	0.12
		0009637 (Response to blue light)	0.24
		0009640 (Photomorphogenesis)	0.08
		0009909 (Regulation of flower development)	0.0002
		0006338 (Chromatin remodeling)	0.00005
		0046283 (Antocyanin metabolic process)	0.08
		0010118 (Stomatal movement)	0.12
II	Q9XHD8_SOLLC	0007623 (Circadian rhythm)	0.006
		0046777 (Protein amino acid autophosphorylation)	0.34
		0006281 (DNA repair)	0.0002
		0006118 (Transport)	0.08
		0009785 (Blue light signaling pathway)	0.34
		0009414 (Response to water deprivation)	0.12
		0009637 (Response to blue light)	0.24
		0009640 (Photomorphogenesis)	0.08
		0009909 (Regulation of flower development)	0.0002
		0006338 (Chromatin remodeling)	0.00005
		0046283 (Antocyanin metabolic process)	0.08
		0010118 (Stomatal movement)	0.12
II	Q93VS0_SOLLC	0007623 (Circadian rhythm)	0.007
		0046777 (Protein amino acid autophosphorylation)	0.31
		0006281 (DNA repair)	0.0002
		0006118 (Transport)	0.08
		0009785 (Blue light signaling pathway)	0.31
		0009414 (Response to water deprivation)	0.12
		0009637 (Response to blue light)	0.23
		0009640 (Photomorphogenesis)	0.08
		0009909 (Regulation of flower development)	0.0003
		0006338 (Chromatin remodeling)	0.00006
		0046283 (Antocyanin metabolic process)	0.08
		0010118 (Stomatal movement)	0.12
II	Q6YBV9_PEA	0007623 (Circadian rhythm)	0.03
		0046777 (Protein amino acid autophosphorylation)	0.28
		0006281 (DNA repair)	0.02
		0006118 (Transport)	0.11
		0009785 (Blue light signaling pathway)	0.28
		0009414 (Response to water deprivation)	0.15
		0009637 (Response to blue light)	0.25
		0009640 (Photomorphogenesis)	0.11
		0009909 (Regulation of flower development)	0.02
		0006338 (Chromatin remodeling)	0.02
		0046283 (Antocyanin metabolic process)	0.11
		0010118 (Stomatal movement)	0.15
II	Q6EAN1_PEA	0007623 (Circadian rhythm)	0.03
		0046777 (Protein amino acid autophosphorylation)	0.28
		0006281 (DNA repair)	0.02
		0006118 (Transport)	0.11

Subgroup	***Protein name***	***Predicted biological process GO terms***	***Posterior probability***
		0009785 (Blue light signaling pathway) 0009414 (Response to water deprivation) 0009637 (Response to blue light) 0009640 (Photomorphogenesis) 0009909 (Regulation of flower development) 0006338 (Chromatin remodeling) 0046283 (Antocyanin metabolic process) 0010118 (Stomatal movement)	0.28 0.15 0.25 0.11 0.02 0.02 0.11 0.15
II	AC174468_14.1	0007623 (Circadian rhythm) 0046777 (Protein amino acid autophosphorylation) 0006281 (DNA repair) 0006118 (Transport) 0009785 (Blue light signaling pathway) 0009414 (Response to water deprivation) 0009637 (Response to blue light) 0009640 (Photomorphogenesis) 0009909 (Regulation of flower development) 0006338 (Chromatin remodeling) 0046283 (Antocyanin metabolic process) 0010118 (Stomatal movement)	0.02 0.27 0.007 0.10 0.27 0.14 0.24 0.10 0.007 0.007 0.10 0.14
II	Q0GKU4_BRACM	0007623 (Circadian rhythm) 0046777 (Protein amino acid autophosphorylation) 0006281 (DNA repair) 0006118 (Transport) 0009785 (Blue light signaling pathway) 0009414 (Response to water deprivation) 0009637 (Response to blue light) 0009640 (Photomorphogenesis) 0009909 (Regulation of flower development) 0006338 (Chromatin remodeling) 0046283 (Antocyanin metabolic process) 0010118 (Stomatal movement)	0.04 0.23 0.03 0.11 0.23 0.15 0.22 0.11 0.03 0.03 0.11 0.15
II	Q1JU52_BRANA	0007623 (Circadian rhythm) 0046777 (Protein amino acid autophosphorylation) 0006281 (DNA repair) 0006118 (Transport) 0009785 (Blue light signaling pathway) 0009414 (Response to water deprivation) 0009637 (Response to blue light) 0009640 (Photomorphogenesis) 0009909 (Regulation of flower development) 0006338 (Chromatin remodeling) 0046283 (Antocyanin metabolic process) 0010118 (Stomatal movement)	0.04 0.23 0.03 0.11 0.23 0.15 0.22 0.11 0.03 0.03 0.11 0.15
II	NP_567341.1	0007623 (Circadian rhythm) 0046777 (Protein amino acid autophosphorylation) 0006281 (DNA repair) 0006118 (Transport) 0009785 (Blue light signaling pathway) 0009414 (Response to water deprivation) 0009637 (Response to blue light) 0009640 (Photomorphogenesis)	0.02 0.35 0.00000001 0.19 0.35 0.24 0.33 0.19

Subgroup	***Protein name***	***Predicted biological process GO terms***	***Posterior probability***
		0009909 (Regulation of flower development) 0006338 (Chromatin remodeling) 0046283 (Antocyanin metabolic process) 0010118 (Stomatal movement)	0.00000002 0.000000002 0.19 0.24
II	A7NUY5_VITVI	0007623 (Circadian rhythm) 0046777 (Protein amino acid autophosphorylation) 0006281 (DNA repair) 0006118 (Transport) 0009785 (Blue light signaling pathway) 0009414 (Response to water deprivation) 0009637 (Response to blue light) 0009640 (Photomorphogenesis) 0009909 (Regulation of flower development) 0006338 (Chromatin remodeling) 0046283 (Antocyanin metabolic process) 0010118 (Stomatal movement)	0.007 0.22 0.0002 0.07 0.22 0.10 0.18 0.07 0.0003 0.00006 0.07 0.10
II	NP_001052950.1	0007623 (Circadian rhythm) 0046777 (Protein amino acid autophosphorylation) 0006281 (DNA repair) 0006118 (Transport) 0009785 (Blue light signaling pathway) 0009414 (Response to water deprivation) 0009637 (Response to blue light) 0009640 (Photomorphogenesis) 0009909 (Regulation of flower development) 0006338 (Chromatin remodeling) 0046283 (Antocyanin metabolic process) 0010118 (Stomatal movement)	0.07 0.26 0.07 0.12 0.26 0.14 0.21 0.12 0.08 0.07 0.12 0.14
II	NP_001047200.1	0007623 (Circadian rhythm) 0046777 (Protein amino acid autophosphorylation) 0006281 (DNA repair) 0006118 (Transport) 0009785 (Blue light signaling pathway) 0009414 (Response to water deprivation) 0009637 (Response to blue light) 0009640 (Photomorphogenesis) 0009909 (Regulation of flower development) 0006338 (Chromatin remodeling) 0046283 (Antocyanin metabolic process) 0010118 (Stomatal movement)	0.07 0.26 0.07 0.12 0.26 0.14 0.21 0.12 0.08 0.07 0.12 0.14
III	AC122161_5.2	0007623 (Circadian rhythm) 0046777 (Protein amino acid autophosphorylation) 0006281 (DNA repair) 0006118 (Transport) 0009785 (Blue light signaling pathway) 0009414 (Response to water deprivation) 0009637 (Response to blue light) 0009640 (Photomorphogenesis) 0009909 (Regulation of flower development) 0006338 (Chromatin remodeling) 0046283 (Antocyanin metabolic process) 0010118 (Stomatal movement)	0.6 0.09 0.08 0.06 0.09 0.12 0.16 0.06 0.29 0.08 0.06 0.12

Subgroup	*Protein name*	*Predicted biological process GO terms*	*Posterior probability*
III	AC122171_26.2	0007623 (Circadian rhythm) 0046777 (Protein amino acid autophosphorylation) 0006281 (DNA repair) 0006118 (Transport) 0009785 (Blue light signaling pathway) 0009414 (Response to water deprivation) 0009637 (Response to blue light) 0009640 (Photomorphogenesis) 0009909 (Regulation of flower development) 0006338 (Chromatin remodeling) 0046283 (Antocyanin metabolic process) 0010118 (Stomatal movement)	0.06 0.09 0.08 0.06 0.09 0.12 0.16 0.06 0.29 0.08 0.06 0.12
III	NP_171935.1	0007623 (Circadian rhythm) 0046777 (Protein amino acid autophosphorylation) 0006281 (DNA repair) 0006118 (Transport) 0009785 (Blue light signaling pathway) 0009414 (Response to water deprivation) 0009637 (Response to blue light) 0009640 (Photomorphogenesis) 0009909 (Regulation of flower development) 0006338 (Chromatin remodeling) 0046283 (Antocyanin metabolic process) 0010118 (Stomatal movement)	0.000000001 0.000000001 0.000000001 0.000000001 0.000000001 0.43 0.53 0.000000001 0.76 0.32 0.000000001 0.43
III	NP_849588.1	0007623 (Circadian rhythm) 0046777 (Protein amino acid autophosphorylation) 0006281 (DNA repair) 0006118 (Transport) 0009785 (Blue light signaling pathway) 0009414 (Response to water deprivation) 0009637 (Response to blue light) 0009640 (Photomorphogenesis) 0009909 (Regulation of flower development) 0006338 (Chromatin remodeling) 0046283 (Antocyanin metabolic process) 0010118 (Stomatal movement)	0.000000002 0.000000002 0.000000002 0.000000002 0.000000002 0.43 0.53 0.000000002 0.76 0.32 0.000000002 0.43

Table 17: Biological process GO term predictions made by SIFTER-X for blue-light photoreceptor proteins. Subgroup II are Cryptochrome1 and orthologous genes, and subgroup III are Cryptochrome2 and orthologous genes. This table was taken from Jöcker et al. 2009.

b) Prediction accuracy comparison for molecular function GO terms between SIFTER & SIFTER-X

I tested SIFTER and SIFTER-X on a test set of 232 *A. thaliana* genes and before running the applications (see chapter VI2e) I removed all ontology term annotations (GO terms, EC numbers, MapMan bins, KO terms) to the tested gene. SIFTER-X showed an increased sensitivity compared to SIFTER at all tested cutoffs. The specificity was increased at all cutoffs greater than 0.2 (see the table 15). At a cutoff of 0.5 the sensitivity of SIFTER-X was increased by 11% (from 44% to 55%) and the specificity by 5% (from 91% to 96%) compared to SIFTER (see table 18). Both SIFTER and SIFTER-X showed a better specificity than transferring GO terms by a Blast search at all cutoffs (see ROC plot in figure 19). However, a very high sensitivity can be reached by transferring molecular function GO terms from Blast hits at the expense of a very low specificity (smaller than 50%).

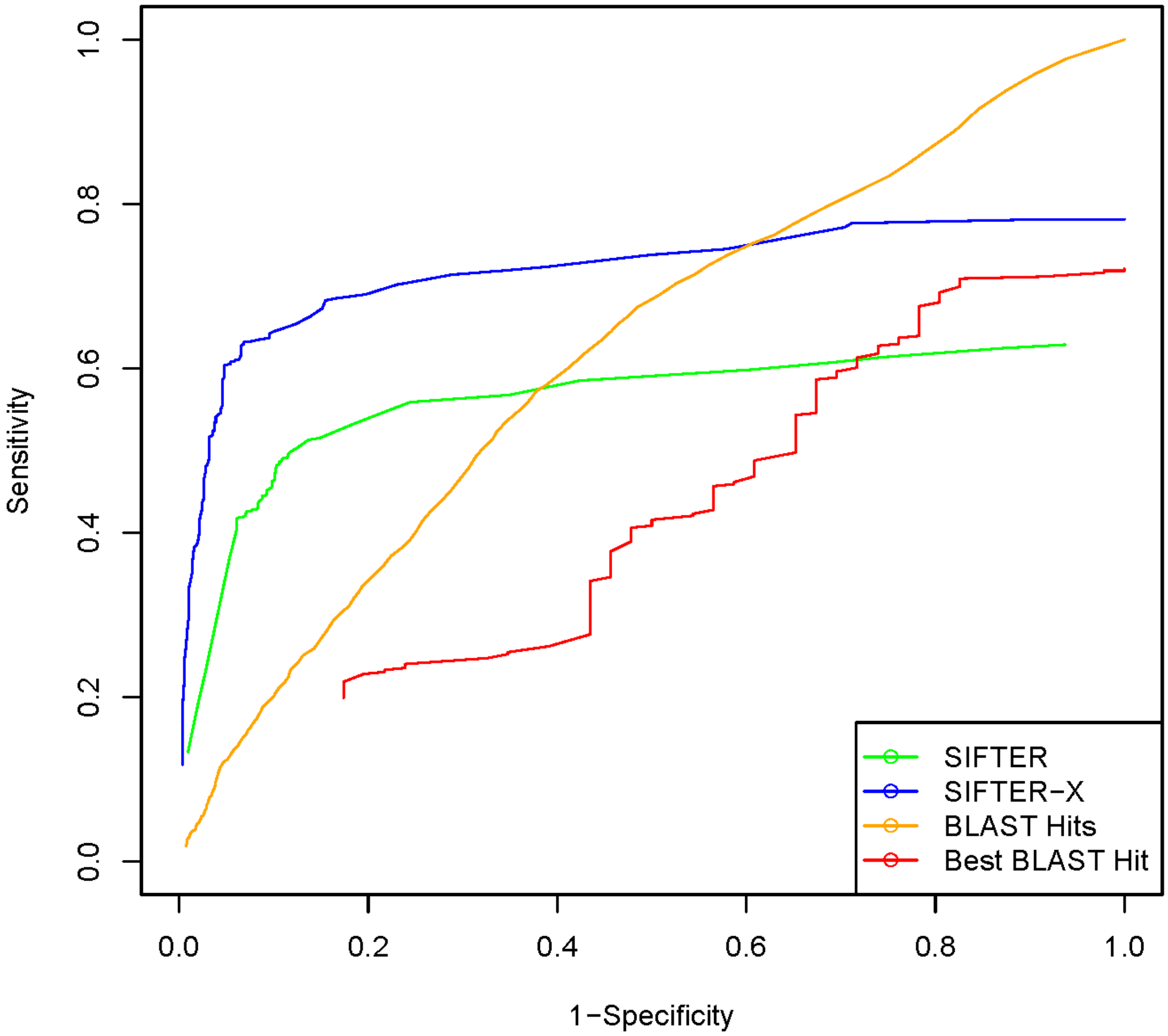


Figure 18: Sensitivity and Specificity of SIFTER, SIFTER-X, Blast hits and the best Blast hit for different cutoffs. While for SIFTER and SIFTER-X the posterior probability is used as cutoff, in case of Blast the e-value is taken. This figure was taken from Jöcker et al. 2009.

Posterior probability cutoff	***Sensitivity SIFTER***	***Sensitivity SIFTER-X***	***Specificity SIFTER***	***Specificity SIFTER-X***
0.1	0.52	0.70	0.85	0.76
0.2	0.48	0.64	0.90	0.90
0.3	0.46	0.61	0.90	0.94
0.4	0.45	0.59	0.91	0.95
0.5	0.44	0.55	0.91	0.96
0.6	0.44	0.52	0.92	0.96
0.7	0.43	0.51	0.92	0.97
0.8	0.43	0.46	0.92	0.97
0.9	0.42	0.36	0.93	0.99
>=1.0	0,13	0.12	0.99	0.996

Table 18: Comparison between the sensitivity and specificity of SIFTER and SIFTER-X on a test set of 232 Arabidopsis genes. All values are rounded. This table was taken from Jöcker et al. 2009.

c) KEGG ontology, MapMan bin and EC term prediction accuracy

SIFTER-X was further tested on the prediction of KO terms, MapMan bins and EC numbers using the same test set. As shown in figure 19 SIFTER-X achieved a very high sensitivity and specificity at all tested posterior probability cutoffs for the prediction of MapMan bins, KO terms and EC numbers.

The average sensitivity of SIFTER-X when predicting MapMan bins and KO terms is about 80% with an average specificity of about 88%. This result can be further increased by using a cutoff of 0.8 for the posterior probability (Sensitivity: 81%, Specificity: 93,5%) (see table 19). For the prediction of EC numbers an average sensitivity of 78% could be achieved at an average specificity of 65%. The specificity is not as high as for MapMan bin and KO term prediction, but it can be significantly increased by choosing a posterior cutoff of 0.9. At this cutoff the sensitivity is about 77% and the specificity is about 82% (see table 19). SIFTER-X can predict MapMan bins, KO terms and EC numbers at a better sensitivity than GO terms.

	Sensitivity			**Specificity**		
Posterior probability cutoff	**KO term prediction**	**MapMan bin prediction**	**EC number prediction**	**KO term prediction**	**MapMan bin prediction**	**EC number prediction**
0.1	0.90	0.92	0.92	0.72	0.71	0.34
0.2	0.89	0.89	0.89	0.83	0.82	0.51
0.3	0.88	0.88	0.89	0.88	0.87	0.59
0.4	0.88	0.86	0.89	0.89	0.88	0.62
0.5	0.86	0.85	0.88	0.89	0.90	0.64
0.6	0.85	0.84	0.87	0.90	0.90	0.68
0.7	0.83	0.84	0.87	0.91	0.92	0.72
0.8	0.81	0.81	0.86	0.93	0.94	0.74

	Sensitivity			**Specificity**		
0.9	0.73	0.77	0.77	0.95	0.96	0.82
1.0	0.35	0.37	0.34	0.97	0.98	0.89

Table 19: The sensitivity and specificity of SIFTER-X in the prediction of KO terms, MapMan bins and EC numbers on a test set of 232 Arabidopsis genes. This table is taken from Jöcker et al. 2009.

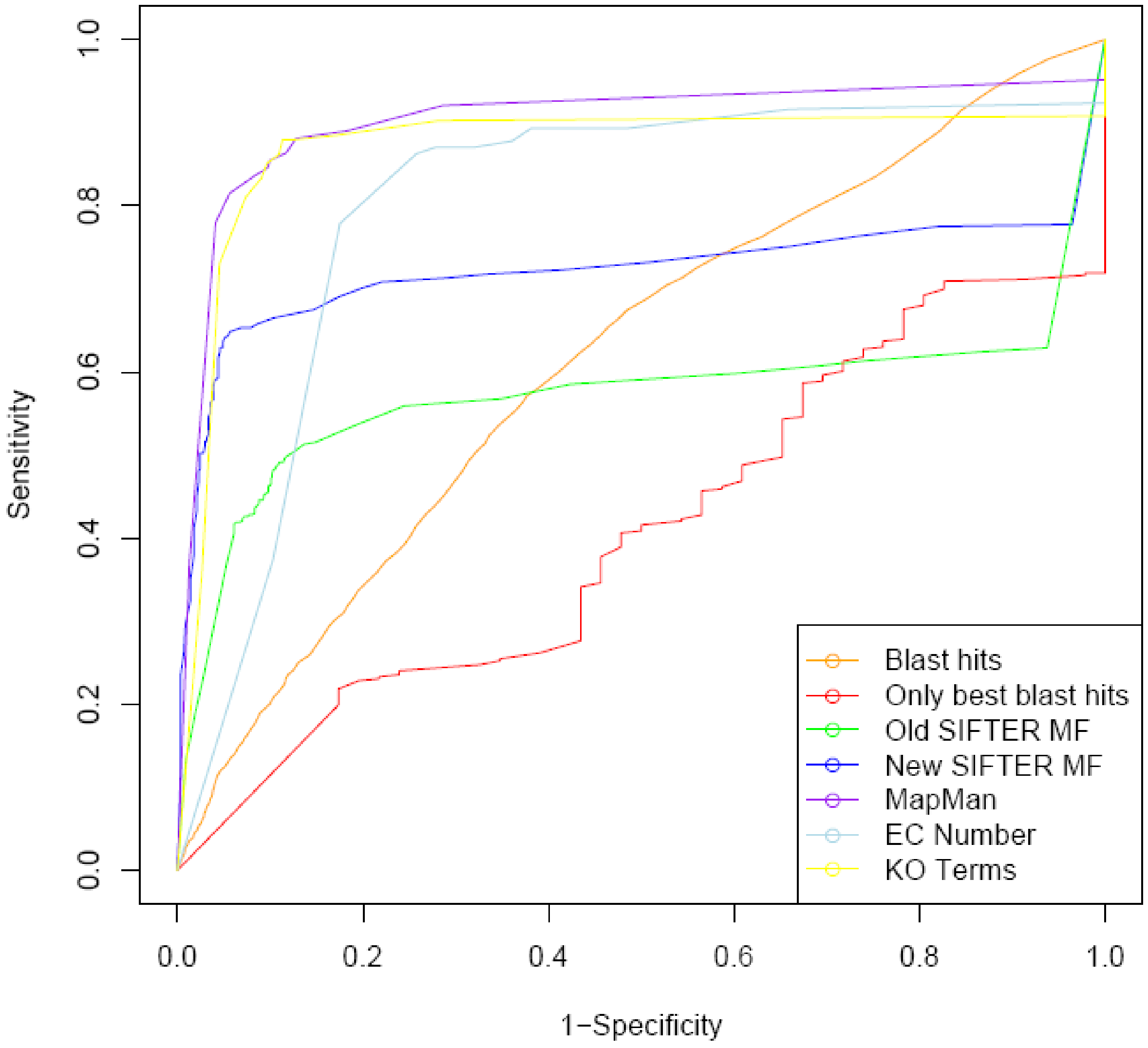


Figure 19: ROC plot for different functional ontologies. For comparison SIFTER and Blast predictions for molecular function (MF) GO terms are included. This figure is taken from Jöcker et al. 2009.

4) Discussion

A new phylogenomics tool SIFTER-X for the automatic function prediction of different ontology terms has been introduced and tested on the blue-light photoreceptor/photolyase family and on a test set of 232 curated *A. thaliana* genes. SIFTER-X builds on the SIFTER algorithm [Engelhardt et al. 2005] and uses additional functional attributes available for genes in the phylogenetic tree to calculate a functional mutation rate which is used to either slow down mutation (in case of same attributes) or speed up mutation (in case of different attributes) within the SIFTER-X framework. Besides the prediction of molecular function GO terms SIFTER-X is able to predict GO biological process GO terms, MapMan bins, KO terms and EC numbers.

I have shown that SIFTER-X is able to predict molecular function GO terms and biological process GO terms in case of the blue-light photoreceptor/photolyase family very accurately. SIFTER-X was able to differentiate between blue-light photoreceptor proteins and photolyase proteins and assigned the true molecular function GO term GO:0003904 (deoxyribodipyrimidine photo-lyase activity) and the true biological process GO term GO:0006281 (DNA repair) to all photolyase proteins with a very high posterior probability to all photolyase proteins. Only one protein (NP_015031.1 from *Saccharomyces cerevisiae*) got a lower posterior probability of 0.73 for the true biological process GO term GO:0006281 (DNA repair). All putative blue-light photoreceptor proteins got the true GO terms with the best posterior probabilities annotated. SIFTER in comparison was not able to differentiate between the families and assigned the wrong GO term GO:0003904 with the best posterior probability to 6 of 16 putative blue light photoreceptors and with the second best posterior probability to four putative blue light photoreceptors. This result might be due to the sparse molecular function GO terms annotated to blue-light photoreceptor genes (Only Cryptochome1 and Cryptochome2 of *A. thaliana* are functionally characterized with ontology terms) and maybe because all functions of only one of the cryptochomes are experimentally proven.

SIFTER-X predicted for 10 proteins GO term GO:0009882 (blue light photoreceptor activity) with a probability smaller than 0.2. This could be a problem if a higher posterior probability cutoff is applied on all results, because then the GO term would be a false negative. The problem here could be that there are only two proteins annotated with this GO term and only one of them has the GO term annotated with an evidence code, which indicates that the function is experimentally verified. This result could be further improved by excluding the biological process GO terms for the functional mutation rate prediction, because there are only few common biological process GO terms annotated to both functionally characterized blue-light photoreceptor genes (Cryptochome1 and Cryptochome2 of *A. thaliana*). One reason for that could be that some biological process GO terms are not annotated for Cryptochome2 of *A. thaliana* yet. In addition to that Cryptochome1 and Cryptochome2 seem to have the same molecular function but are involved in overlapping but also different biological processes (sub-functionalization) [Ahmad et al. 1998]. On the basis of this example it might be a good idea to use biological process GO terms only for the prediction of biological process GO terms and to exclude them from the calculation in other cases. However, this is only one example for which this method would increase the accuracy. Further test sets are needed to find out if that is often the case. Another idea to overcome this problem is to change the weighting of the biological process GO terms for the calculation of the functional mutation rate.

For the prediction of biological process GO terms SIFTER-X was further able to differentiate between Cryptochrome1 and Cryptochrome2 subgroups of proteins and assigned GO terms which are annotated to Arabidopsis Cryptochrome1 and Cryptochrome2 to proteins in both subgroups and GO terms, which are just annotated to Cryptochrome1 or Cryptochome2 to only proteins in the corresponding subgroup. However, there is no evidence for the putative orthologous genes of Cryptochrome2 in Medicago that the function is the same as annotated to the Arabidopsis Cryptochrome2. Based on the assumption that this is the case, SIFTER-X is in this example able to handle neo-functionalization and sub-functionalization, which often occur after gene duplication events [Hurles. 2004] [Presgraves 2005]. To proof this result further test sets are needed.

However the posterior probability of GO terms predicted for orthologs of Cryptochrome1 and Cryptochrome2, which are experimentally verified for Cryptochrome1 and Cryptochrome 2 genes of *A. thaliana*, is very low. This could lead to many false negatives after applying a higher probability cutoff. But the probability of the true GO terms is much higher than for the wrong GO terms, so it would make sense to use an adaptive partitioning process instead of using a fixed cutoff for the biological process output in the SIFTER-X framework. For example the partitioning could be done by using the maximal distance between two probabilities in an ordered list of all probabilities and their corresponding GO terms. Only these GO terms are then printed out, which are higher than this “individualized” cutoff.

Tested on a manually and experimentally curated dataset of 232 *A. thaliana* genes SIFTER-X was able to predict molecular function GO terms with a sensitivity of 55% and a very high specificity of 96%. This is an increase of 11% (from 44% to 55%) sensitivity and an increase of 5% (from 91% to 96%) specificity in comparison with the SIFTER algorithm. This means that SIFTER-X is able to predict more molecular function GO terms and gives a more accurate representation of the function of the protein by having a lower false prediction rate.

Furthermore both SIFTER and SIFTER-X achieved a better accuracy than transferring the molecular function GO terms of the best Blast hit. However by transferring the function of Blast hits a better sensitivity could be achieved, but at the cost of a specificity lower than 50%. At this specificity the annotated GO term can be true or wrong. The better sensitivity could be obtained, because no overlap cutoff was applied as it was done in the homologous search included in the phylogenomic pipeline (see chapter V2a). Maybe some functionally related genes are excluded by this approach. Using an alternative method like a Hidden-Markov Model search [Karplus et al. 1998] using an alignment of Blast hits with a low overlap cutoff or a profile search using PSIBlast [Altschul and Koonin 1998] [Repsys et al. 2008] or RPSBlast [Marchler-Bauer et al. 2002] for the homolog detection may further increase the sensitivity, but these approaches would be more time-consuming.

I have also shown that SIFTER-X is able to predict MapMan bins, KO terms and EC numbers with a very high accuracy. By applying a posterior probability cutoff of 0.8 for the MapMan bin and KO term predictions, the sensitivity of SIFTER-X is about 81% with a specificity between 93% and 94%. For the prediction of EC numbers a sensitivity of 77% and a specificity of 82% could be achieved by using a higher cutoff of 0.9. The sensitivity for the prediction of MapMan bins, KO terms and EC numbers is better in comparison to GO term prediction, because in our test set only few terms are annotated to one gene. In case of GO often many different terms have to be annotated to one gene to describe the full function of the protein, however for many proteins the set of annotated GO terms is still incomplete [Kourmpetis et al. 2007].

The SIFTER-X results may further be improved by integrating structure data. Unfortunately this kind of information is only available for few proteins and a fast structure comparison tool would be needed to compare the structure of different proteins. But instead of comparing 3D structures, 2D structure data could be used or critical residues could be identified from the alignment by using a structure of one protein as template [Ng and Henikoff 2006]. In addition to that interaction data, already being used to compute the functional mutation rate, interaction partners could be compared between different species to identify "orthologous interactions". For this approach a database which includes orthologous relationships between genes is required, because finding out if two interaction partners are orthologous or not would be too time-consuming.

VII. Manual curation in genome projects

1) Introduction

After the automatic annotation process in genome projects is complete, a manual curation of the functional annotation is necessary, because functional annotation tools have limitations and at the moment no tool performs equally well for all kind of genes [Godzik et al. 2007]. By manual curation, wrong functional annotations can be corrected and the function of a gene can be further specified. The curation step can be done by a group of curators who manually compare the results from different function prediction programs and annotate the right function to the corresponding gene afterwards. This approach has the advantage that if no tool is able to return a significant result the comparison between results from different prediction programs or intermediate results from an analysis workflow can give clues to the function of a gene [Friedberg 2006]. However, this step is very time-consuming, because each tool has its own scores and trusted cutoffs and one has to switch between different web pages to compare different analysis results, which are often in different formats. Furthermore for some programs no web page is available and the installation and execution of the program is often difficult. In the latter case web services (see chapter III2) can help, because they offer interoperability, they enable an easy data retrieval, they use standardized formats and they do not require to install anything on ones local computer or to have a huge amount of resources available [Neerincx and Leunissen 2005].

This motivates to implement an automatic functional annotation system (called AFAWE – Automatic Functional Annotation in a distributed Web service Environment) suitable for any organism and any kind of protein coding gene in an easily extensible client-server architecture with an intuitive web interface that facilitates a fast comparison between analysis results. All functional prediction tools are run as web services and web service workflows to enable the fast integration of new tools and workflows. The results are displayed in a graphically and tabularly manner and trustworthy results of each analysis are highlighted to enable an easier and faster comparison between the results and to address the problem of comparing different results at several webpages. Each user of AFAWE is able to add his/her own functional annotation to each gene by assigning ontology terms (like GO, KO, MapMan, ...), pathways, in which the gene is involved and/or add a human readable description line. AFAWE is used in the Medicago genome project as well as in the tomato genome project to encourage biologists and other scientists to add manual functional annotations to as many genes as possible. To update the former automatic functional annotation in both genome projects, AFAWE is connected to the sequence database MIPSPlantsDB [Spannagl et al. 2007] via an AFAWE web service, which provides all manual annotations available for each gene.

This chapter describes the development and application of the AFAWE system. I introduce and test different web services to find out if they are suitable for AFAWE and I will describe how these web services and additional programs have been wrapped as BioMOBY web services [The BioMOBY consortium, 2008] to register them at a central repository and to use standardized input and output datatypes. Additionally I explain how AFAWE was designed and implemented and how these web services and web service workflows have been integrated into the AFAWE system architecture. Furthermore the integration of functional information from AFAWE in MIPSPlantsDB is explained.

In the result section an example is given how a manual annotation can be done by using AFAWE.

2) Material & Methods

a) Finding suitable web services for the AFAWE system

For the automatic function prediction suitable web services and web services for functional data retrieval were searched in the literature and by asking people at different institutes.

Of all web services found, analysis and data retrieval web services from the ToolBus software [Eckart and Sobral 2003] provided at the Virginia Bioinformatics Institute (VBI), from the European Bioinformatics Institute (EBI) [Labarga et al. 2007] and from the National Center for Biotechnology Information (NCBI) [Sayers et al. 2008] were tested (see table 20).

To implement a client program, which is able to call each web service, Java classes were automatically generated by using the WSDL file[15] and the wsdl2Java program from the APACHE Axis1.2 library[16]. These Java classes were used in self-written Java web service client programs to set the inputs for the web services, run them and get the outputs.

Provider	***Web service***	***Data/Analysis***
VBI	Phylip	Phylogenetic tree construction via programs from the PHYLIP package [Felsenstein 1993]
VBI	PDBj	Data retrieval from the structure database PDBj [Standley et al. 2008]
EBI	InterProScan	Protein domain prediction via InterProScan [Mulder and Apweiler 2007]
EBI	WU-Blast	Searching for homologous sequences by using WU-Blast [Gish 1996-2004]
EBI	DBFetch	Data retrieval from all databases hosted at the EBI (e.g. UniProt [The UniProt Consortium 2007], InterPro [Mulder and Apweiler 2007])
NCBI	ESearch	Searches and returns primary IDs (for use in EFetch and ESummary) and term translations for a given ID or keyword.
NCBI	EFetch	Retrieves database records for a given primary ID or a list of primary IDs
NCBI	EGQuery	Returns number of database records for a specific keyword
NCBI	ESummary	Returns document summaries for a list of primary IDs

Table 20: Tested web services from the VBI, EBI and NCBI.

Afterwards the tested EBI web services (InterProScan, WU-Blast and DBFetch) were wrapped as BioMoby web services. To provide additional web services for the automatic functional analysis of genes, the RPS-Blast search (Version 2.2.13) against the Conserved Domain Database (CDD) [Marchler-Bauer et al. 2007] and a NCBI Blast search [Altschul et al. 1997] against the manually built RefSeq database from chapter V2a have been implemented as BioMoby web services too.

All web services were registered at the main BioMoby repository in Canada (http://biomoby.org/mobycentral/) and their input and output datatypes were defined (see table 21). With the help of the BioMoby dashboard [The BioMoby consortium 2008] Java classes for the service implementation were automatically generated and were used as superclass for the Java class implementations of the web services. All EBI web services are run inside the BioMoby web service

[15] *http://www.w3.org/TR/wsdl*

[16] *http://ws.apache.org/axis/*

via system call of the corresponding client program provided by EBI to enable a faster replacement of the client program if the EBI web service is being updated. The RPS-Blast and the NCBI Blast program are also run as system call, but are executed by using the lsrun command from the LSF batch system[17] to run them on a compute cluster. The web service implementations were deployed together with all BioMoby libraries on the JBoss application server [Fleury and Reverbel 2003] using the APACHE Axis 1.3 library[18] to make them publicly accessible.

EBI web service / Program	***Name of BioMoby web service***	***Input Datatype: Name***	***Additional attributes Datatype: Name***	***Output Datatype: Name***
InterProScan	EBI_InterproScan	AminoAcidSequence : sequence MobyObject:email	String: apps (Applications to run) String: outformat (Output format) String: seqtype (Protein or DNA)	text-plain: interproscan_result
WU-Blast	EBI_WU_Blast	AminoAcidSequence : sequence MobyObject:email	Float: e_threshold (E-value threshold) String: program (Blast program to run) String: database (Database to use) String: outformat (Output format) Integer: numal (Number of alignments to show in result)	text-plain: Blast_result
RPS-Blast against CDD	RPSBlast	AminoAcidSequence : sequence	String: seqtype (Protein or DNA) String: low_complexity_filter (Which filer should be used) Float: e_threshold (E-value threshold)	text-plain: rpsBlast_result
NCBI Blast against manual RefSeq database	Blast_Against_RefSeq _Complete_Sequenced _Organisms	AminoAcidSequence : sequence	String:output (output format) Integer: numberOfDescriptions (Number of description lines to show in result) Integer: numberOfAlignments (Number of alignments to show in result) String: filterHits (Which filer should be used) Float: e_threshold (E-value threshold)	text-plain: Blast_result

Table 21: EBI web services and other programs were wrapped as BioMoby web services. The additional parameters are settings for the underlying program and have default values.

[17] *http://www.platform.com/Products/platform-lsf*

[18] *http://ws.apache.org/axis/index.html*

b) The AFAWE design

AFAWE is implemented in a flexible J2EE structure to make it easily extensible and platform independent (see figure 20 and 21). A MySQL database[19] helps to avoid bottlenecks in data retrieval over the Internet, in case that former analysis results for a given protein sequence are available. Data Access Objects[20] and Data Transfer Objects[21] are used to get analysis results from the database and to store the results in the database. An intuitive web frontend is responsible for the interaction between user and the program. The AFAWE core application forms the middleware between database and web interface. It receives the user input, controls all web services and workflows, parses the results and uses the Data Access Objects and Data Transfer Objects to store the analysis results in the database.

After starting AFAWE the user can choose between retrieving former analysis results via an AFAWE internal protein ID, search for proteins by any keyword and starting a new automatic functional prediction.

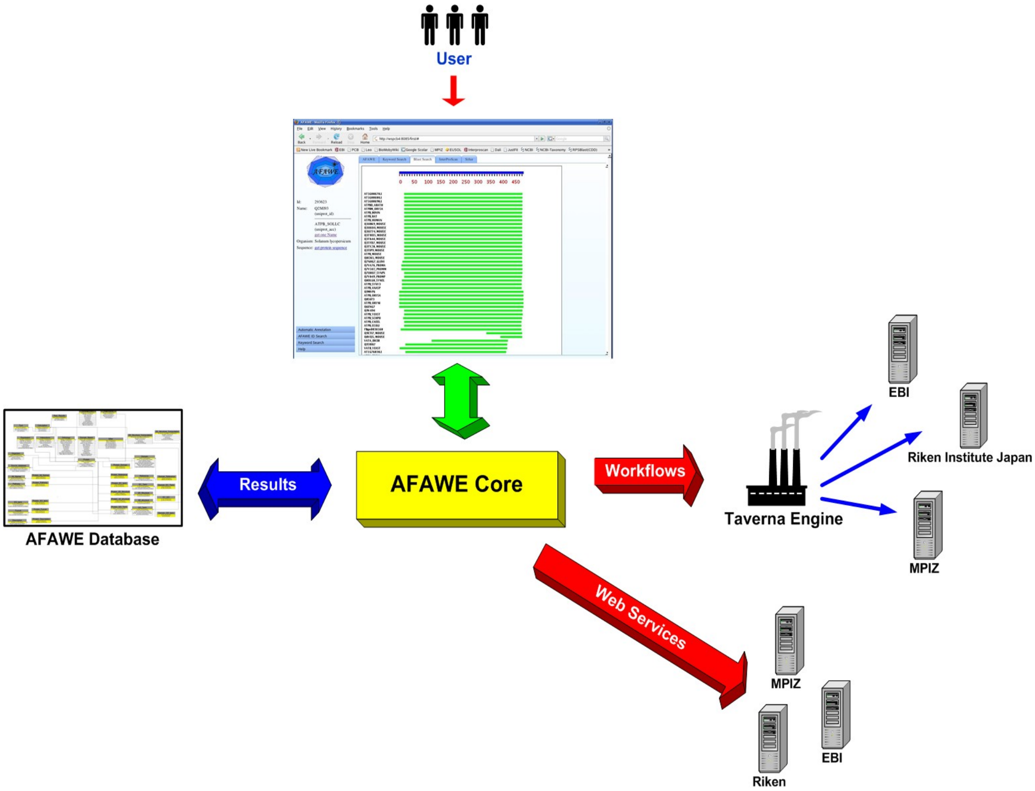


Figure 20: Three layer structure in AFAWE.

If the user has selected the automatic functional annotation and has chosen the necessary analysis tools, web services and workflows are run in parallel. Whereas for running the web services BioMOBY web service clients are used, workflows are run by the Taverna workflow engine. If results are available, they are parsed, stored in the database and immediately displayed in the user web frontend. Using a cache database gives the user the possibility to view the results for the protein whenever he or she likes, without running the analyses again. The results are deleted, if newer results become available (for example if a database has been updated).

[19] *http://www.mysql.com/*
[20] *http://en.wikipedia.org/wiki/Data_Access_Object*
[21] *http://en.wikipedia.org/wiki/Data_Transfer_Object*

To enable a faster comparison of the results, trustworthy results of each analysis are highlighted by applying different filters on the result data (see the following sub-chapter d). Furthermore the user is able to add a manual annotation to each protein after logging into the AFAWE system. Besides different ontology terms like GO terms, FunCats and KEGG ontology terms, also a human readable description, name of pathways and references can be added. Each ontology term has to be verified by adding the corresponding evidence code to it. Besides adding new annotations, the user is also able to negate existing annotations (e.g. The user may say, that a former automatically assigned function has been proven to be wrong).

The manual annotation is afterwards visible to every other user, even if the user is not logged in, and can be used to improve functional annotations in different genome projects.

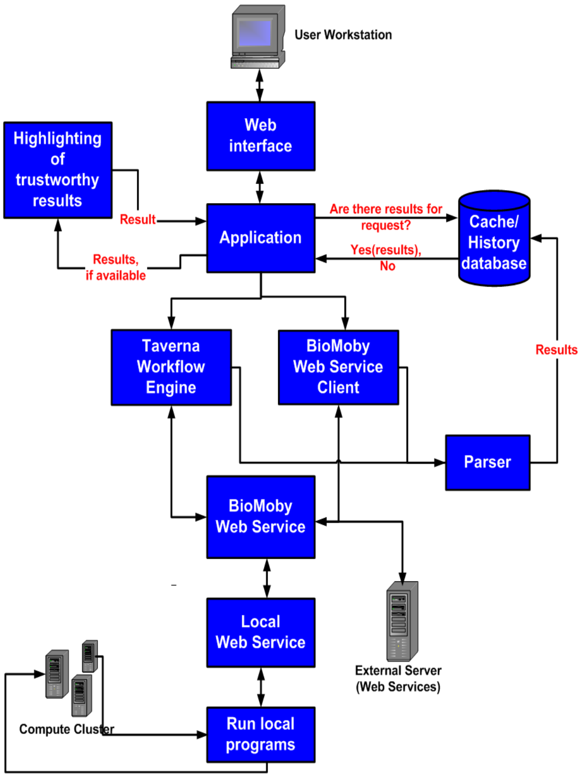


Figure 21: AFAWE application overview.

c) *Analyses*

There are analyses available for homolog detection, protein domain search and function prediction by using phylogenomics. For the homolog detection the BioMoby wrapped EBI WU-Blast web service (see chapter VII2a) is run against both the UniProt database [The UniProt Consortium 2007] and its manually verified part, the SwissProt database. If detected homologous proteins are not already stored in the database, the EBI DBFetch web service [Labarga et al 2007] is called to get additional information about the protein like assigned GO terms, EC numbers, protein domains, synonyms and the sequence from the UniProt database. Additionally a NCBI protein Blast web service, hosted at the Max-Planck Institute for Plant Breeding Research, is called to run against the manually built RefSeq database introduced in chapter V2a. Proteins from this database with corresponding GO terms from the Gene Ontology website had already been stored in the AFAWE database (see chapter V2a).

Protein domains are discovered by the BioMoby wrapped EBI InterProScan web service (see chapter VII2a) and by a web service, which runs a RPS-Blast search against the Conserved Domain Database (CDD) at the NCBI (see chapter VII2a). To provide the user also with an automatic annotation of GO terms the new phylogenomic workflow with SIFTER (see chapter V2b) is run by using the Taverna API.

d) AFAWE Filter

As mentioned before we implemented dynamic filters to highlight trustworthy hits from the different analysis results to enable a faster comparison of the results. In the following, filters for the different analyses, available in AFAWE, are described.

EBI WU-Blast Filter

To filter out putative homologous proteins from the Blast result of the EBI WU-Blast web service, we provide five filters. One applies an overlap cutoff to all Blast hits, so that only hits, which have more than 70% overlap between query and hit sequence are highlighted in the result table. The second filter considers protein domain composition. This filter highlights hits, which have the same domains as the query sequence. Protein domains for the hits are retrieved from the UniProt database [The UniProt Consortium 2007] by using the EBI DBFetch web service. Protein domains assigned to proteins in the UniProt database have been predicted by InterProScan and therefore the filter compares them with predicted protein domains of the query using the same tool. Domains that are not listed in UniProt are ignored. Furthermore PROSITE domains are excluded, because the PROSITE pattern search uses regular expressions to detect conserved domains and therefore does not assign any score, making it harder to detect false positives.

SIFTER filter

For SIFTER a simple threshold filter is used. All GO terms, which have got a probability assigned by SIFTER of 0.4 and higher are highlighted. The cutoff of 0.4 was chosen, because by evaluating 100 genes of Sorghum for wrong assigned GO terms it was revealed, that GO terms with a probability greater than 0.4 were true in 97% of all cases (see chapter V).

e) Implementation of the AFAWE database

We implemented the AFAWE database as a MySQL database (MySQL 5.0.18), because MySQL is a free, fast and reliable relational database[22].

Each protein in the database is well-defined by its sequence and its corresponding organism. Information about the protein itself (e.g. ontology terms and protein domains obtained from diverse databases, 3D and 2D structure, alternative names (synonyms and database identifier) and references) and analysis results for this protein are separated. Analyses results are divided into the categories “Interaction”, “Orthology”, “Structure” and “Sifter”. Each of these analysis database tables include a foreign key to the “Calculation” table, in which metadata for the analysis, like the date and time when the analysis was run and the tool used, are stored. Furthermore an additional database table provides the information whether for a protein an analysis was run before and if there are analysis results available (see figure 22). To enable the user to login into AFAWE and add his or her own functional annotation, also a “User” table is added, in which important information about the user, like name, affiliation and research interest, can be stored and which is connected to all protein information tables (see figure 22).

[22] *http://dev.mysql.com/doc/refman/5.0/en/index.html*

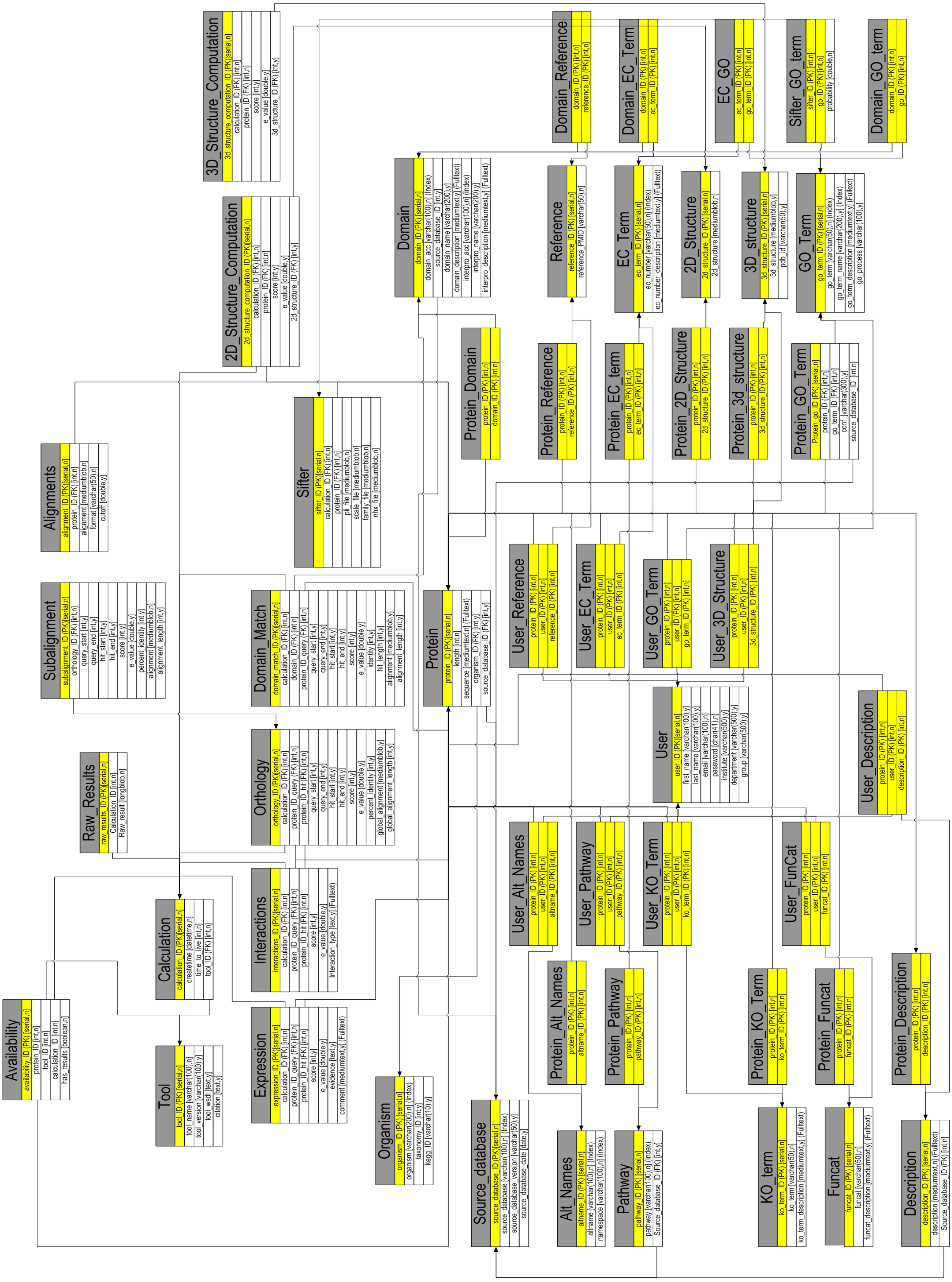


Figure 22: AFAWE database table schema. Primary keys are colored in yellow. Arrows between table columns indicate foreign keys. Each table column is specified by its data type and a Y or N, which denotes if NULL is allowed or not as entry.

To encourage project members of the International Medicago Genome Annotation Group (IMGAG), the International Tomato Genome Annotation Group ITAG and other researcher to improve the automatic functional annotation of the Medicago and tomato proteins, all available proteins from *Medicago truncatula* and 9942 tomato protein sequences (batch11 of the ITAG pipeline) with analysis results from SIFTER are included.

f) Data Access Objects and Transfer Objects

For storing and getting data from the AFAWE database Data Access Objects (DAOs) and Data Transfer Objects (TFs) are implemented in a first version by Martin Kocent during his practical student-ship and extended by myself afterwards. DAOs include SQL statements, which are executed via the Java JDBC API (version 1.5)[23]. Transfer objects are used to store the results from database queries or to transfer data from the middleware into the database.

For each table one TF, which includes 'get' and 'set' methods for all table columns, and one DAO is provided. Both, TFs and DAOs are implemented for easy extensibility, which means further database tables and the corresponding TFs and DAOs can easily be added. All TFs and DAOs are stored in a separated Java archive file (jar file) to make them usable for web services and other projects.

g) Integration of the Taverna workflow engine

To run workflows in AFAWE the WorkflowLauncherWrapper, which is part of the Taverna API [Oinn et al. 2004] is used. Because the WorkflowLauncherWrapper uses Raven[24] to define dependencies of Java libraries and update them regularly via the Internet, it was necessary to build a separate Java archive (RunWorkflows.jar) to avoid this update mechanism, which is not possible in the deployed AFAWE system. Besides the Taverna workflow execution, RunWorkflows.jar parses also the results of the SIFTER pipeline and stores them afterwards in the AFAWE database.

This Java archive can also be used independently of the AFAWE web interface to run the SIFTER pipeline without using the AFAWE API (e.g. if SIFTER should be run for a batch of proteins)

h) Development of the AFAWE web interface

The AFAWE web interface was designed by myself and implemented by Fabian Hoffmann and Andreas Jöcker. The DOJO Java Script framework[25] in combination with the Java Servlet technology was used, because of its ability to build web interfaces with filtered tables and navigation tabs in a short time. In addition the AJAX technology [Holdener et al 2008] was included to increase the interactivity, speed, functionality and usability of the web page.

For each analysis there is one HTML page and one Java Servlet implemented. Whereas the HTML page includes the AJAX and DOJO elements and is used for displaying of data, the Java Servlets are used for adding dynamic content to the HTML pages.

i) Connection to the MIPSPlantsDB database

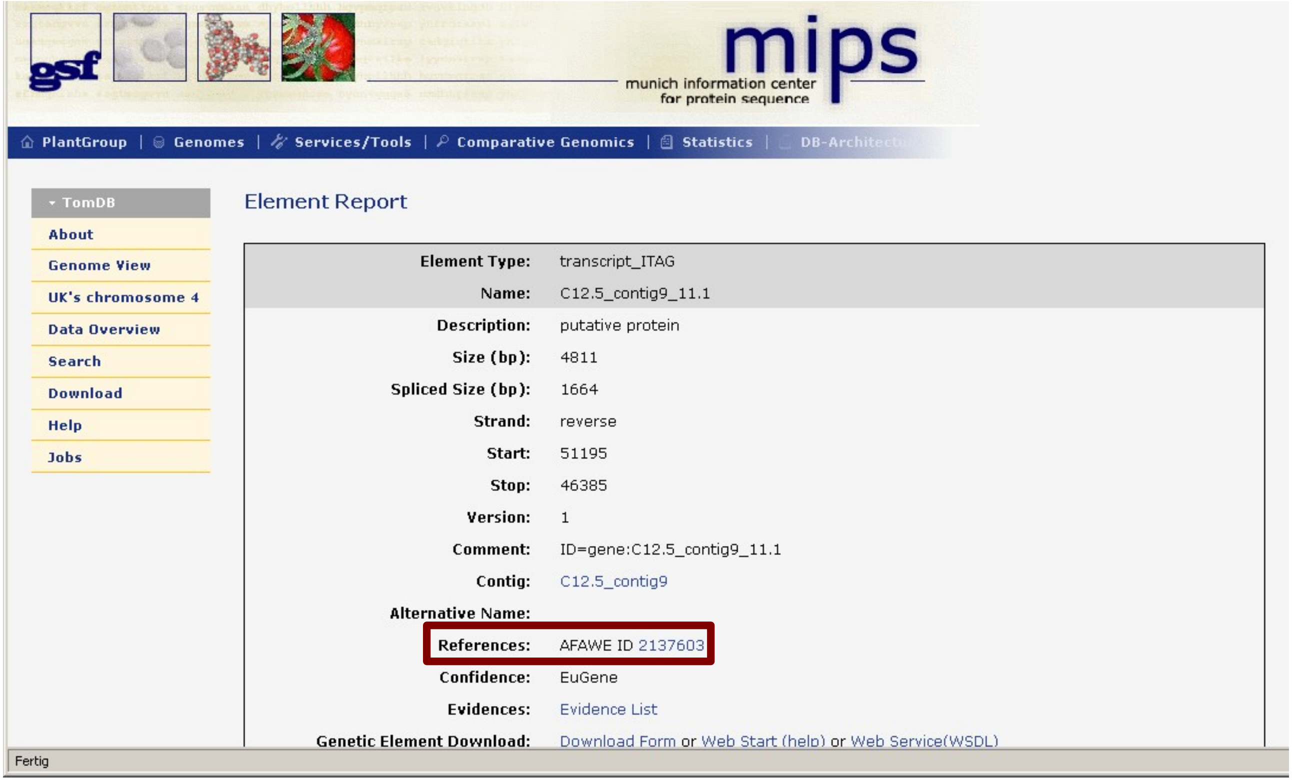
To provide a link between the MIPSPlantsDB and AFAWE and so to enable users of MIPSPlantsDB

[23] *http://Java.sun.com/j2se/1.5.0/docs/api/*

[24] *http://raven.rubyforge.org/*

[25] *http://dojotoolkit.org/*

to add their manual annotation in AFAWE, primary protein identifiers (primary ID of the Protein table (see figure 22)) of all Medicago and tomato proteins, extracted from the AFAWE database, were integrated as a cross-reference in the MIPSPlantsDB by Manual Spannagl and a link was created on the element report website to go directly to the corresponding analysis results in AFAWE (see figure 23).


gsf
mips
munich information center for protein sequence
PlantGroup | Genomes | Services/Tools | Comparative Genomics | Statistics | DB-Architecture
TomDB
About
Genome View
UK's chromosome 4
Data Overview
Search
Download
Help
Jobs
Fertig

Element Report

Element Type:	transcript_ITAG
Name:	C12.5_contig9_11.1
Description:	putative protein
Size (bp):	4811
Spliced Size (bp):	1664
Strand:	reverse
Start:	51195
Stop:	46385
Version:	1
Comment:	ID=gene:C12.5_contig9_11.1
Contig:	C12.5_contig9
Alternative Name:	
References:	AFAWE ID 2137603
Confidence:	EuGene
Evidences:	Evidence List
Genetic Element Download:	Download Form or Web Start (help) or Web Service(WSDL)

Figure 23: Element Report of the tomato gene C12.5_contig9_11.1 in MIPSPlantsDB with a cross-reference to the corresponding analysis results in AFAWE.

Furthermore to display manual annotations added by AFAWE users at the element report webpage of MIPSPlantsDB, several BioMoby web services [The BioMoby Consortium 2008] were implemented to retrieve data from the AFAWE database and to start remotely AFAWE analysis tools (see table 22). The getAutomaticAndManualAnnotationByAFAWE_ID web service should be called interactively on the MIPSPlantsDB element report website to get the up-to-date manual annotation, as well as the automatically predicted functional annotation from the AFAWE database and to display this information on the site. If there is no AFAWE ID stored for the corresponding protein in MIPSPlantsDB, the getAFAWEProteinIDBySequenceAndOrganism web service can be called to retrieve the AFAWE ID from the AFAWE database.

Name of web service	*Description*
getAFAWEProteinIDByGOTerm	Returns all AFAWE IDs, which have the given GO term or a child of this GO term annotated. By using the secondary parameters of the web service the organism can be chosen and if all AFAWE IDs should be returns or just AFAWE_IDs, which have the corresponding GO term experimentally verified or curated. Furthermore it can be set if only automatically derived GO terms should be considered, manually annotated GO terms or all terms.
getAFAWEProteinIDBySequenceAndOrganism	Returns the AFAWE ID stored for the given amino acid sequence and organism.
getAutomaticAndManualAnnotationByAFAWE_ID	Returns manually added functional information and automatically predicted functions for a given AFAWE ID
runAFAWEAnalysesBySequenceAndOrganismAndGetAFAWE_URL	Starts all AFAWE analysis tools using the given amino acid sequence and the corresponding organism as input and returns an AFAWE URL to the AFAWE analysis results. As secondary parameter the user can set which analysis tools should be run.

Table 22: AFAWE web services to retrieve data from the AFAWE database and run remotely AFAWE analysis tools.

3) Results

a) Finding suitable web services for the AFAWE system

Web services provided at the European Bioinformatics institute (EBI) [Labarga et al. 2007], the National Center for Biotechnology Information (NCBI) [Sayers et al. 2008] and the Virginia Bioinformatics Institute (VBI) [Eckart and Sobral 2003] were found suitable for the automatic functional annotation and for up-to-date protein data information retrieval. All these institutes offer client programs for download, which can be used to interact with the web service.

The VBI has supported its own client program ToolBus for calling its web services and visualizing the results. For some web services a license, a so called AAA ticket, is necessary, as for some of the provided web services a fee is charged or they are only free for the academical use. Due to the missing documentation of the web service and the strong connection between the web service client ToolBus and the VBI web services, using the VBI web services without the ToolBus client is difficult. Also, these web services are only available via the ports 6565 and 7575 and for using the web services these ports have to be open, which is normally not the case. Because of these difficulties, I decided not to integrate these services into the AFAWE system.

On the contrary the EBI provides separate clients for their web services and no specific port has to be open to use them. All clients are available in different programming languages (Perl, Java, C#) and with sufficient documentation how to use them. For resource reasons some web services (e.g. InterProScan and DBFetch) are restricted in the number of inputs. In case of InterProScan only one sequence is allowed as input and for DBFetch a maximal number of 200 database accessions can be given as input. All web services are faster than using the EBI web frontend, but running

InterProScan with more than one query is much faster than running InterProScan with each sequence individually. If the analysis is very time-consuming, the asynchronous mode can be chosen instead of waiting for the results in synchronous mode. In this mode the user gets back a job ID, which can be used later to fetch the results. The results can be retrieved in different output formats. Unfortunately these output formats and the whole web service has changed a lot in the last two years (2006-2008) and so parsers to get specific fields from the entries have to be often re-written. Additionally without the documentation it is very difficult for the user to find out what kind of input the web service supports and which of the inputs are mandatory.

The latter is also true for the NCBI web services, but the NCBI provides five coupled web services to access all kinds of data in their databases. In contrast to the DBFetch web service, which is provided from the EBI to get entries from EBI databases, NCBI supports the access to specific fields inside an entry (e.g. GO terms). This approach is very fast and no parser is needed to retrieve these specific fields. But this complex structure makes a comprehensive documentation extensive and so only examples are shown in the documentation. Unfortunately are these examples not sufficient. Another drawback is, that the NCBI offers only retrieval web services and no analysis web services.

I decided to use for the AFAWE system web services from the EBI and the NCBI, because they provide a good support for their web services (e.g. provide a mailing list), their data seems to be always up to date and the web services are dependable.

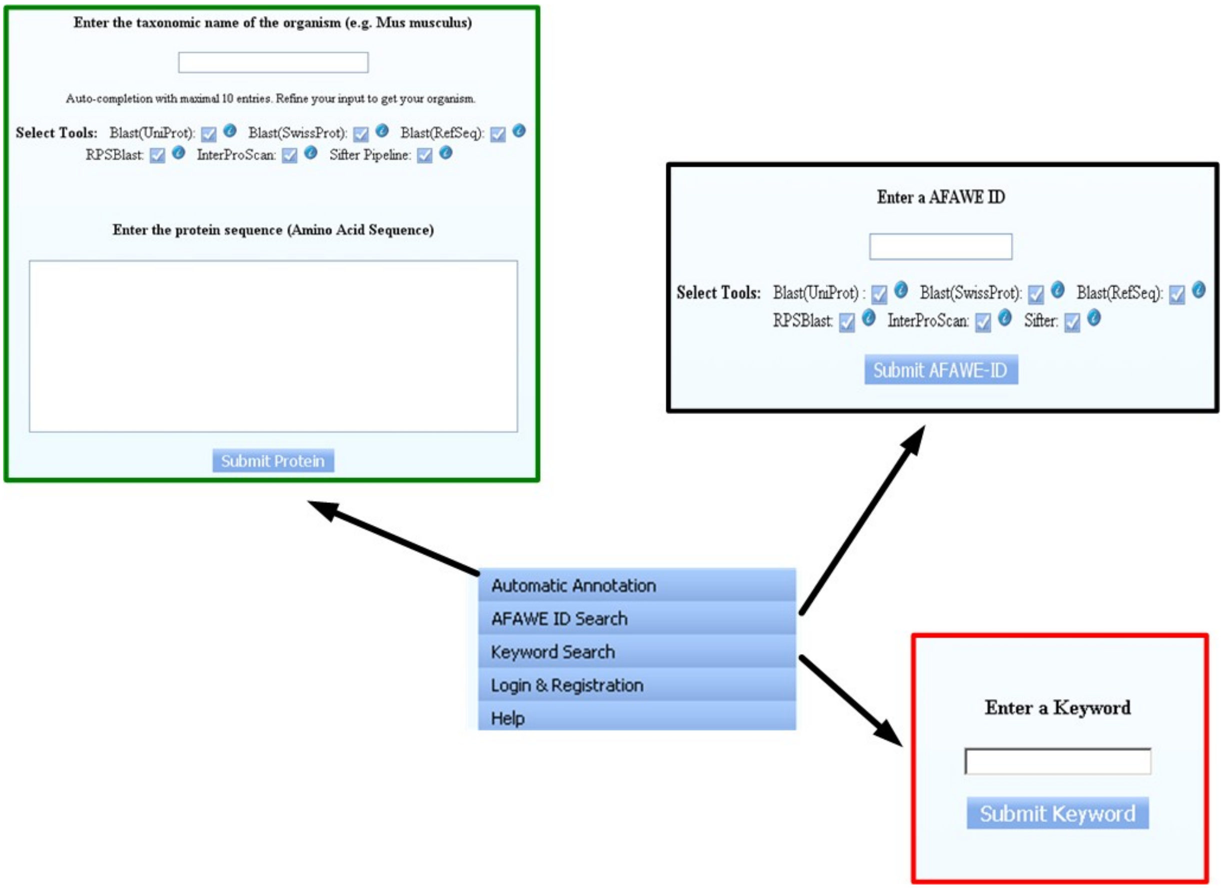
To enable a fast and easy integration of web services in AFAWE, all web services, which were found suitable for the manual functional annotation, were standardized by semantically defining their inputs and outputs and by registering them at a central repository. Therefore the EBI web services InterProScan and WU-Blast were wrapped as BioMoby web services [The BioMoby Consortium 2008]. The NCBI web service was too complex to wrap, so it is used together with the DBFetch web service inside the AFAWE application for getting database entries for genes/proteins, for which no functional or sequence information is available.

Two additional programs (RPS-Blast against the CDD database and NCBI Blast against the manually build database introduced in chapter V2a) were implemented as BioMoby web services and are also publicly available at the Max-Planck Institute for Plant Breeding Research. The underlying databases are regularly updated.

b) How to do a manual annotation using AFAWE

To show how a manual functional annotation can be done using the AFAWE system, I will use the *Medicago truncatula* gene AC144389_35.2 as an example (The example was taken from Jöcker et al. 2008).

By searching gene AC144389_35.2 using the keyword search (see figure 24), four different analysis results (Blast against UniProt and SwissProt, InterProScan and SIFTER) are available. Each analysis result is displayed in a different tab, available at the upper site of the browser window (see figure 25). The phylogenomic pipeline with SIFTER is the more reliable analysis in comparison with Blast, because it also takes duplication and speciation events in account. SIFTER has predicted three different GO terms ("electron transporter, transferring electrons within CoQH2-cytochrome c reductase complex activity" (GO:0045153), "stearoyl-CoA 9-desaturase" (GO:0004768) and "enzyme activator activity" (GO:0008047)). GO term GO:0045153, which has an assigned posterior probability of ~0.98, is highlighted and therefore the most reliable.



Enter the taxonomic name of the organism (e.g. Mus musculus)

Auto-completion with maximal 10 entries. Refine your input to get your organism.

Select Tools: Blast(UniProt): ☑ Blast(SwissProt): ☑ Blast(RefSeq): ☑ RPSBlast: ☑ InterProScan: ☑ Sifter Pipeline: ☑

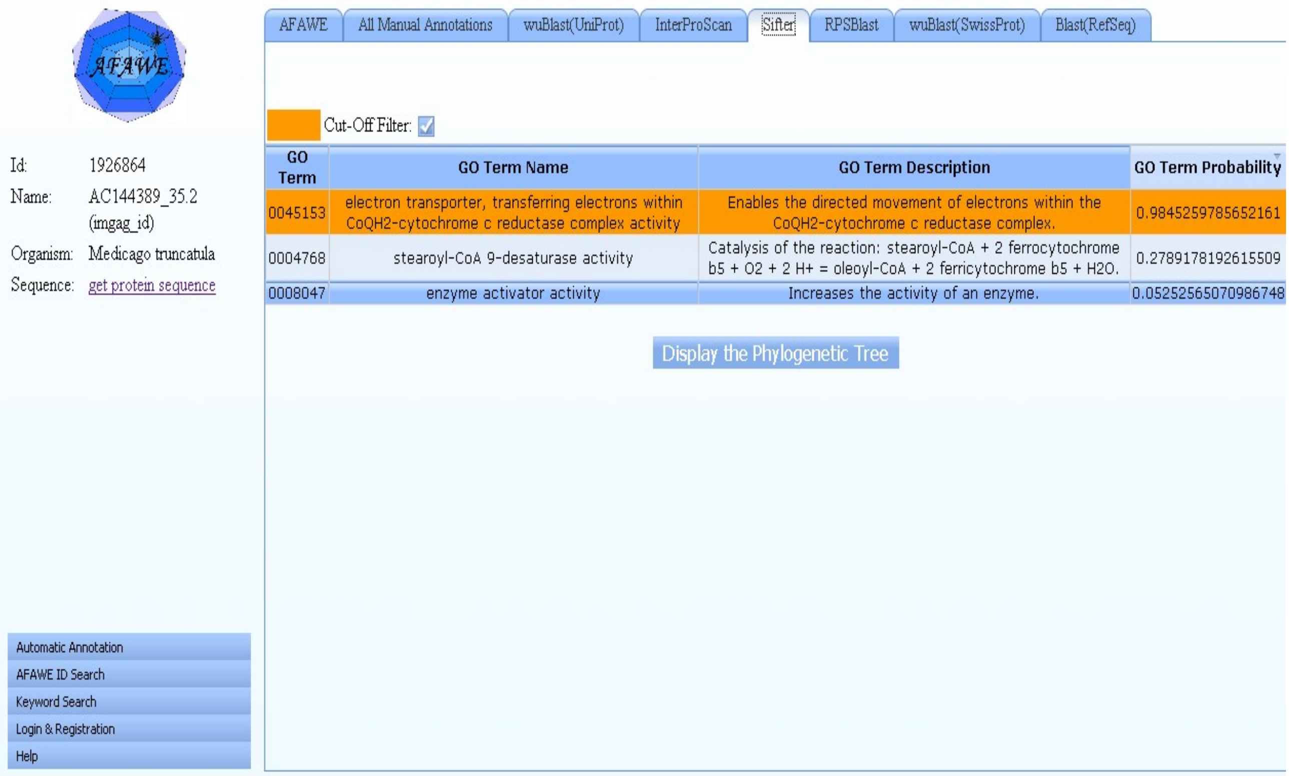
Enter the protein sequence (Amino Acid Sequence)

Enter a AFAWE ID

Select Tools: Blast(UniProt) : ☑ Blast(SwissProt): ☑ Blast(RefSeq): ☑ RPSBlast: ☑ InterProScan: ☑ Sifter: ☑

Enter a Keyword

Figure 24: There are three ways to get analysis results from AFAWE: By giving an AFAWE ID, entering a keyword or starting a new functional analysis using the "Automatic Annotation" link.

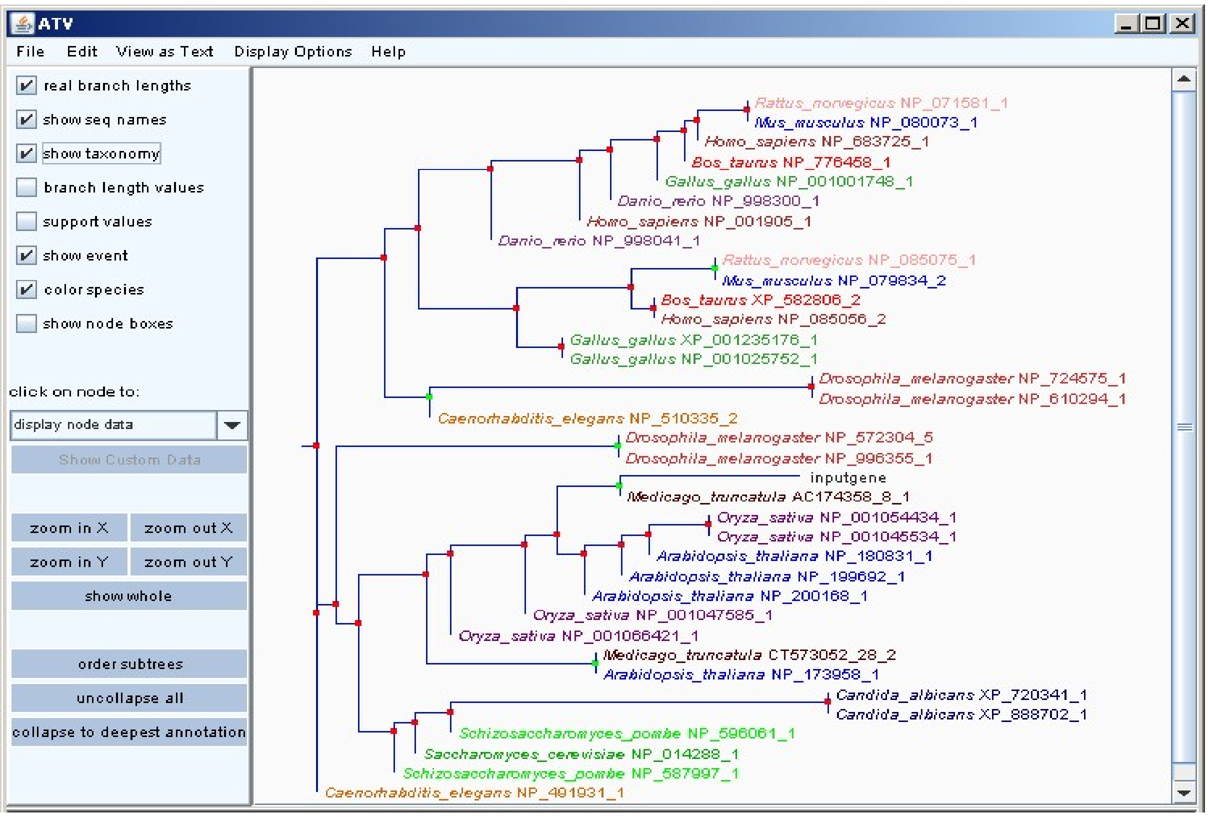


Id: 1926864
Name: AC144389_35.2 (imgag_id)
Organism: Medicago truncatula
Sequence: get protein sequence

GO Term	GO Term Name	GO Term Description	GO Term Probability
0045153	electron transporter, transferring electrons within CoQH2-cytochrome c reductase complex activity	Enables the directed movement of electrons within the CoQH2-cytochrome c reductase complex.	0.9845259785652161
0004768	stearoyl-CoA 9-desaturase activity	Catalysis of the reaction: stearoyl-CoA + 2 ferrocytochrome b5 + O2 + 2 H+ = oleoyl-CoA + 2 ferricytochrome b5 + H2O.	0.2789178192615509
0008047	enzyme activator activity	Increases the activity of an enzyme.	0.05252565070986748

Figure 25: SIFTER result for gene AC144389_35.2 from Medicago truncatula. The best results are highlighted in orange.

By using the “Display the Phylogenetic Tree” button the reconciled phylogenetic tree, which is used as input for SIFTER can be viewed and further investigated (see figure 26).



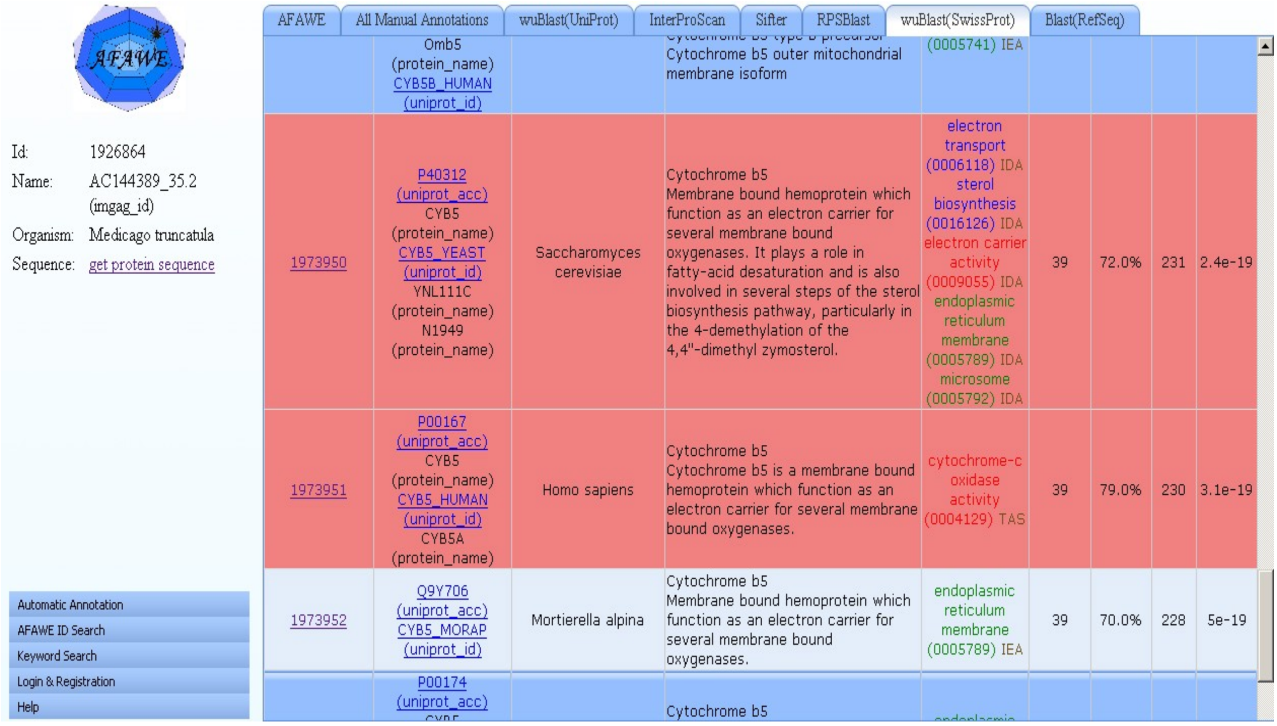
ATV
File
Edit
View as Text
Display Options
Help
Show Custom Data
zoom in X
zoom out X
zoom in Y
zoom out Y
show whole
order subtrees
uncollapse all
collapse to deepest annotation
Rattus_norvegicus NP_071581_1
Mus_musculus NP_080073_1
Homo_sapiens NP_683725_1
Bos_taurus NP_776458_1
Gallus_gallus NP_001001748_1
Danio_rerio NP_998300_1
Homo_sapiens NP_001905_1
Danio_rerio NP_998041_1
Rattus_norvegicus NP_085075_1
Mus_musculus NP_079834_2
Bos_taurus XP_582806_2
Homo_sapiens NP_085056_2
Gallus_gallus XP_001235176_1
Gallus_gallus NP_001025752_1
Drosophila_melanogaster NP_724575_1
Drosophila_melanogaster NP_610294_1
Caenorhabditis_elegans NP_510335_2
Drosophila_melanogaster NP_572304_5
Drosophila_melanogaster NP_996355_1
inputgene
Medicago_truncatula AC174358_8_1
Oryza_sativa NP_001054434_1
Oryza_sativa NP_001045534_1
Arabidopsis_thaliana NP_180831_1
Arabidopsis_thaliana NP_199692_1
Arabidopsis_thaliana NP_200168_1
Oryza_sativa NP_001047585_1
Oryza_sativa NP_001066421_1
Medicago_truncatula CT573052_28_2
Arabidopsis_thaliana NP_173958_1
Candida_albicans XP_720341_1
Candida_albicans XP_888702_1
Schizosaccharomyces_pombe NP_596061_1
Saccharomyces_cerevisiae NP_014288_1
Schizosaccharomyces_pombe NP_587997_1
Caenorhabditis_elegans NP_491931_1

☑ real branch lengths
☑ show seq names
☑ show taxonomy
□ branch length values
□ support values
☑ show event
☑ color species
□ show node boxes

click on node to: display node data

Figure 26: Reconciled phylogenetic tree of gene AC144389_35.2 from Medicago truncatula and putative homologous genes, which is used as input for SIFTER. The tree is displayed by using the A Tree Viewer Tool (ATV) [Zmasek and Eddy 2001], which is integrated in AFAWE.

By looking at experimentally verified or reviewed molecular function GO terms assigned to Blast hits using the GO term filter (see chapter VII2d), only two proteins (CYB5_YEAST and CYB5_HUMAN) are highlighted in pink (see figure 27). This means that at least one of their assigned molecular function GO terms is experimentally verified or reviewed. Both proteins have more than 70% overlap with the query and share the same domains with the query (both hits are highlighted in yellow, if the overlap and domain filter is switched on) and therefore seem to belong to the same protein family.



Id: 1926864

Name: AC144389_35.2 (imgag_id)

Organism: Medicago truncatula

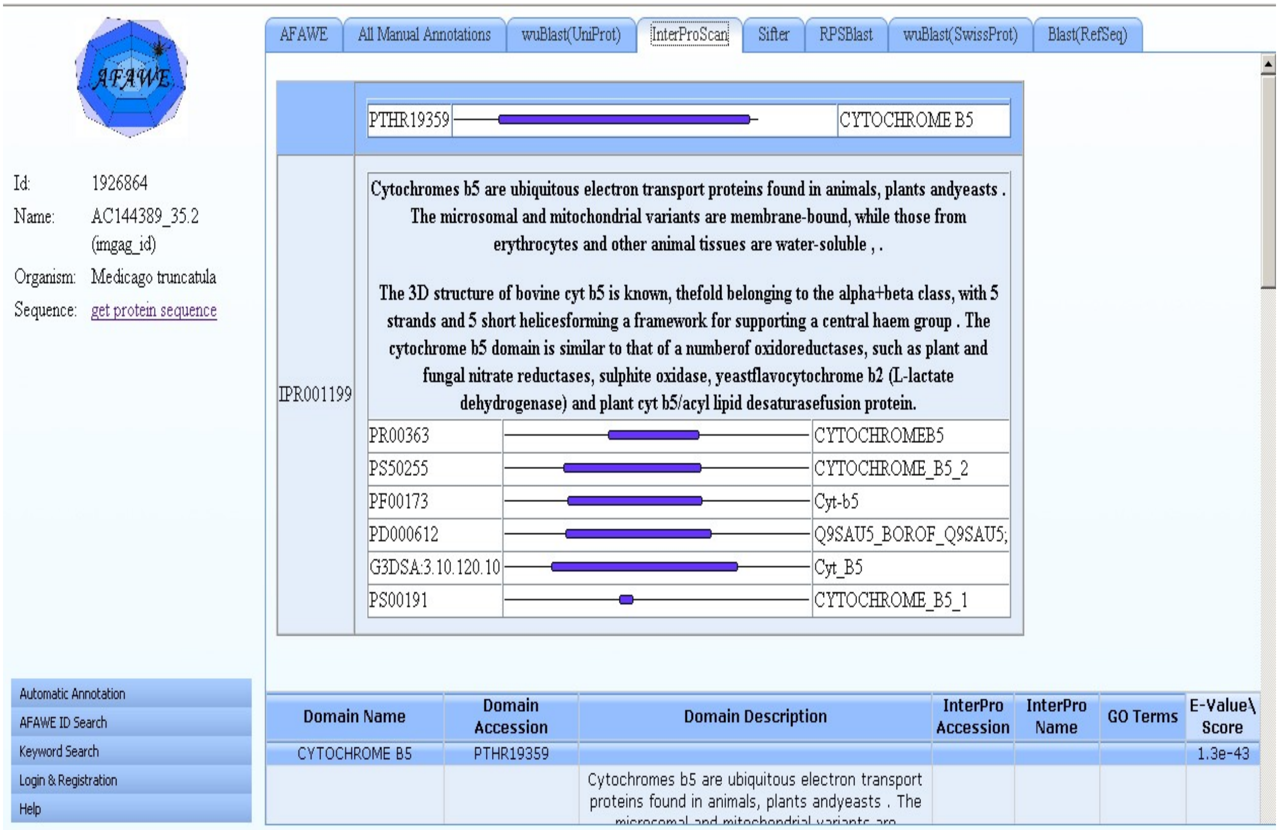
Sequence: get protein sequence

	Omb5 (protein_name) CYB5B_HUMAN (uniprot_id)		Cytochrome b5 type B precursor Cytochrome b5 outer mitochondrial membrane isoform	(0005741) IEA				
1973950	P40312 (uniprot_acc) CYB5 (protein_name) CYB5_YEAST (uniprot_id) YNL111C (protein_name) N1949 (protein_name)	Saccharomyces cerevisiae	Cytochrome b5 Membrane bound hemoprotein which function as an electron carrier for several membrane bound oxygenases. It plays a role in fatty-acid desaturation and is also involved in several steps of the sterol biosynthesis pathway, particularly in the 4-demethylation of the 4,4''-dimethyl zymosterol.	electron transport (0006118) IDA sterol biosynthesis (0016126) IDA electron carrier activity (0009055) IDA endoplasmic reticulum membrane (0005789) IDA microsome (0005792) IDA	39	72.0%	231	2.4e-19
1973951	P00167 (uniprot_acc) CYB5 (protein_name) CYB5_HUMAN (uniprot_id) CYB5A (protein_name)	Homo sapiens	Cytochrome b5 Cytochrome b5 is a membrane bound hemoprotein which function as an electron carrier for several membrane bound oxygenases.	cytochrome-c oxidase activity (0004129) TAS	39	79.0%	230	3.1e-19
1973952	Q9Y706 (uniprot_acc) CYB5_MORAP (uniprot_id)	Mortierella alpina	Cytochrome b5 Membrane bound hemoprotein which function as an electron carrier for several membrane bound oxygenases.	endoplasmic reticulum membrane (0005789) IEA	39	70.0%	228	5e-19
	P00174 (uniprot_acc) CYB5		Cytochrome b5	endoplasmic				

Figure 27: AFAWE analysis results of an EBI WU-Blast search against the SwissProt database and using the Molecular Function filter afterwards. Blast hits (genes), which have an experimentally verified or reviewed molecular function GO term are highlighted in pink.

GO term GO:009055 ("electron carrier activity") assigned to gene CYB5_YEAST from *Saccharomyces cerevisiae* is experimentally verified by direct assay (evidence code "IDA") and is the parent of GO:0045153, which is predicted by the SIFTER pipeline. However, gene CYB5_HUMAN from human has GO term GO:0004129 ("cytochrome-c oxidase activity") assigned by author statement ("TAS") from Proteome Inc., but there is no parent-child relationship to the predicted GO term of SIFTER or the GO term assigned to the yeast gene. To find out which of the GO terms is true and which is wrong, the InterProScan results are investigated.

All protein domains predicted by InterProScan are included in Cytochrome b proteins (see figure 28) and this functional description can also be found in the most description lines of the Blast hits. Cytochrome b is the main subunit of the mitochondrial Cytochrome bc1, which is one of the components of the respiratory chain [Iwata et al. 1998] and is part of the b6f complexes, which is the electronic connection between the photosystem I and the photosystem II of the oxygenic photosynthesis [Kurisu et al. 2003]. In both complexes it is responsible for the transmembrane electron transfer.



Id: 1926864
Name: AC144389_35.2 (imgag_id)
Organism: Medicago truncatula
Sequence: get protein sequence

PTHR19359		CYTOCHROME B5

IPR001199

Cytochromes b5 are ubiquitous electron transport proteins found in animals, plants andyeasts . The microsomal and mitochondrial variants are membrane-bound, while those from erythrocytes and other animal tissues are water-soluble , .

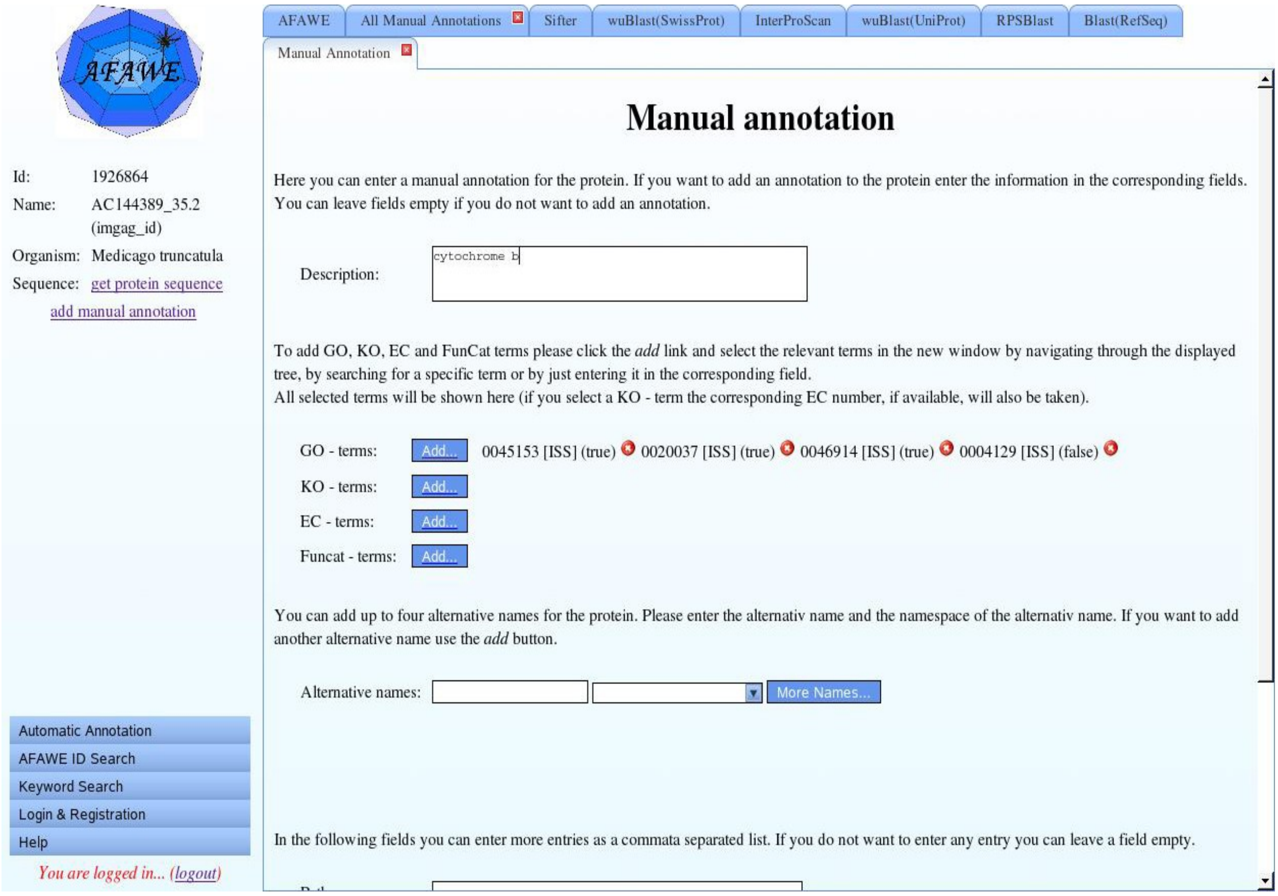
The 3D structure of bovine cyt b5 is known, thefold belonging to the alpha+beta class, with 5 strands and 5 short helicesforming a framework for supporting a central haem group . The cytochrome b5 domain is similar to that of a numberof oxidoreductases, such as plant and fungal nitrate reductases, sulphite oxidase, yeastflavocytochrome b2 (L-lactate dehydrogenase) and plant cyt b5/acyl lipid desaturasefusion protein.

PR00363		CYTOCHROMEB5
PS50255		CYTOCHROME_B5_2
PF00173		Cyt-b5
PD000612		Q9SAU5_BOROF_Q9SAU5;
G3DSA:3.10.120.10		Cyt_B5
PS00191		CYTOCHROME_B5_1

Domain Name	Domain Accession	Domain Description	InterPro Accession	InterPro Name	GO Terms	E-Value\ Score
CYTOCHROME B5	PTHR19359					1.3e-43
		Cytochromes b5 are ubiquitous electron transport proteins found in animals, plants andyeasts . The microsomal and mitochondrial variants are				

Figure 28: AFAWE InterProScan result for Medicago gene AC144389_35.2.

This fits well with GO:0045153, but no evidence could be found for cytochrome-c oxidase activity, which was assigned to the human gene by author statement. Therefore I assume, that this GO term is a wrong annotation and the other one is true.



Id: 1926864
Name: AC144389_35.2 (imgag_id)
Organism: Medicago truncatula
Sequence: get protein sequence
add manual annotation

You are logged in... (logout)

Manual annotation

Here you can enter a manual annotation for the protein. If you want to add an annotation to the protein enter the information in the corresponding fields. You can leave fields empty if you do not want to add an annotation.

Description: cytochrome b

To add GO, KO, EC and FunCat terms please click the *add* link and select the relevant terms in the new window by navigating through the displayed tree, by searching for a specific term or by just entering it in the corresponding field.
All selected terms will be shown here (if you select a KO - term the corresponding EC number, if available, will also be taken).

GO - terms: 0045153 [ISS] (true) 0020037 [ISS] (true) 0046914 [ISS] (true) 0004129 [ISS] (false)
KO - terms:
EC - terms:
Funcat - terms:

You can add up to four alternative names for the protein. Please enter the alternativ name and the namespace of the alternativ name. If you want to add another alternative name use the *add* button.

Alternative names:

In the following fields you can enter more entries as a commata separated list. If you do not want to enter any entry you can leave a field empty.

Figure 29: Manual annotation added to the Medicago gene AC144389_35.2.

To add this manual annotation to the Medicago gene it is necessary to register and login into the AFAWE system. Afterwards a “Add a manual annotation” link is shown below the protein information on the left side of the browser window, which opens the manual annotation window (see figure 29). In this window we can now declare GO:0045153 (“electron transporter”) as true and GO:0004129 (“cytochrome-c oxidase activity”) as false. The added information is afterwards displayed at the “All manual annotation” page.

4) Discussion

We have implemented an easily extensible and intuitive tool for the automatic and manual annotation of any kind of protein coding gene from any kind of organism. Comparison of different analysis results is simplified by using different filters that highlight trustworthy results. AFAWE [Jöcker et al 2008] is publicly available at http://bioinfo.mpiz-koeln.mpg.de/afawe. Several web services have been implemented to retrieve different kind of functional information from the AFAWE database, or to run AFAWE analyses remotely over the internet. This allows an easy integration of AFAWE results into protein report websites of sequence databases. Furthermore AFAWE analysis results are connected via a cross-reference link from MIPSPlantsDB [Spannagl et al 2007]. This will encourage scientists to add their manual annotation in AFAWE, which can be used afterwards to update the former automatically predicted functional annotations and therefore will help to avoid errors in public databases.

However, although some of the web services (RPSBlast and DBFetch) are really fast, the automatic annotation is quite slow, if proteins are not in the database and additional pieces of information have to be retrieved via the EBI DBFetch web service from UniProt or InterPro. To improve the performance, I restricted the number of Blast hits for the search against UniProt to 25 hits and for SwissProt to 20 and all analyses are run in parallel.

Another bottleneck in using web services is, that some of the web services (e.g. the EBI WUBlast and the DBFetch web service) changed their XML output format without announcing it beforehand. Luckily parsers in AFAWE are easily replaceable, but still they have to be updated, if there are any changes in the output format of the service. Furthermore at the moment there is no possibility to check for non-functionality of a web service, except if it returns an error message. Because of that alternative web services should be called, if an analysis web service is not responding at all or in a certain time frame.

VIII. Summary and Discussion

Up to now the functional annotation of genes was done by using homology based sequence similarity searches like Blast. I have shown for different genomes that by using a phylogenomic approach, the function of a gene can be annotated more accurately. Furthermore through the integration of annotated functional information like domain information, interaction data and ontology terms the accuracy could be further improved and the function is described more comprehensively. In addition to that, an intuitive webinterface is being provided which facilitates the comparison between results from different functional analysis tools and enables a manual annotation.

Automatic function prediction in genome projects by function transfer:

I have implemented and tested different approaches for automatic function prediction in the *Medicago truncatula*, *Sorghum bicolor* and *Solanum lycopersicum* genome projects. Because the manual functional characterization of each gene is not possible in genome projects, an automatic pipeline has been implemented. It uses the phylogenomic tool SIFTER for the automatic transfer of molecular function Gene Ontology terms (GO terms) within a phylogenetic tree of homologous genes. Tested on 100 manually verified Medicago proteins, the SIFTER workflow achieved an accuracy of 97% and the assigned functional annotation term was in 25% of the cases more specific than the assigned human readable description line.

However, I also discovered wrong predictions made by SIFTER because there were too sparse molecular function GO terms annotated to genes or the annotated GO term was wrong for one gene in the phylogenetic tree. Another bottleneck of SIFTER was the phylogenetic tree used as input. If this is incomplete, gene loss and duplication nodes can not be discovered.

To get a better alignment and a more comprehensive phylogenetic tree the phylogenomic pipeline was improved in several ways and tested on the *Sorghum bicolor* genome and again on the 100 manually annotated Medicago genes. Although I obtained an increased accuracy of the SIFTER workflow for the 100 manually annotated Medicago genes (100%), still three proteins out of 100 manually annotated Sorghum proteins are wrongly annotated (accuracy: 97%), because of false functional annotations, paralogous genes, which changed their function, and too sparse GO terms assigned to homologous proteins inside the phylogenetic tree. Furthermore I found that it is important to use the low complexity filter integrated in Blast to mask repetitive regions and so lower the number of non-homologous genes and to use a cutoff of 0.4 for the posterior probability predicted by SIFTER to avoid wrong annotations. However, for 55% of all predicted Sorghum genes a molecular function GO term could be assigned. The higher number of annotated genes in the Sorghum genome compared to Medicago could be due to a better gene prediction of the Sorghum genes leading to less truncated and incomplete gene models, because the sequencing, assembly and gene prediction process of the Medicago genome was still in progress at this date.

It might be possible to further increase the number of functionally annotated genes by integrating an additional step in the phylogenomic pipeline after the building of the alignment to get additional, more distant genes. The additional step could be a Hidden Markov Model or profile search using the alignment of the putative homologous genes discovered by the former iterative Blast search as a template to create the profile or the Hidden Markov Model. However, this step would slow down the whole pipeline significantly and so this pipeline would not be applicable on the whole genome anymore.

To further boost the prediction result of the SIFTER pipeline and to overcome problems with paralogous genes and missing GO terms annotated to genes in the phylogenetic tree, the SIFTER algorithm was modified to use additional functional information. Annotated terms from different ontologies, interaction partners and protein domain composition of genes are evaluated to decrease the functional mutation rate between nodes in the tree if they share known functional attributes and increase it if known attributes differ. Furthermore the modified algorithm (called SIFTER-X) was extended to predict besides molecular function GO terms biological process GO terms, EC numbers, MapMan bins and KEGG Ontology terms (KO terms).

I tested the new pipeline using SIFTER-X on the blue-light photoreceptor/photolyase family, which is often wrongly predicted by tools like InterProScan and SIFTER, and on 232 manually verified Arabidopsis genes. SIFTER-X was able to differentiate between the blue-light photoreceptor family and the photolyase family and assigned members of both families the true molecular function and biological process GO terms with a better posterior probability than the wrong GO terms. Furthermore tested on 232 Arabidopsis genes the sensitivity of SIFTER could be increased by 11% (from 44% to 55%) and specificity could be increased by 5% (from 90% to 95%) using SIFTER-X and applying a cutoff of 0.5 for the posterior probability. By predicting the MapMan bins and KO terms SIFTER-X achieved a very high sensitivity of 81% and a high specificity of 93% by using a posterior probability cutoff of 0.8. In case of EC number the cutoff should be increased to 0.9 to get a sensitivity of 77% and a specificity of 82%. This is due to the fact that in case of KO terms, EC numbers and MapMan bins often only one to three terms were annotated to the curated Arabidopsis data set, but in case of GO terms up to 10 terms were necessary to describe the complete function of a gene. However, the set of annotated GO terms is incomplete for many genes [Kourmpetis et al. 2007]. SIFTER-X performed very well on the the prediction of MapMan bins. This maybe an indication of either good reference annotation or the suitability of MapMan bins for automatic classification tasks. In all cases SIFTER-X performed better than transferring the function of genes found by Blast and applying a certain cutoff or transferring the function of the best Blast hit only.

However, in case of members of the blue-light photoreceptor family not all true GO terms got a posterior probability greater than 0.4, because only two blue-light photoreceptor genes were annotated with GO terms. This could lead to a high false negative rate by applying a cutoff of 0.4 to the SIFTER results. The posterior probability could be increased a bit by excluding biological process GO terms for the prediction of molecular function GO terms, because often the molecular function of paralogous genes is equal, but the biological process is different [Duarte et al. 2006], but still then the probability is lower than 0.4 for many true terms. But the distance between the true GO terms with the lowest probability and the wrong GO with the highest probability is large. So this problem could be solved by applying a partitioning cutoff instead of a fixed cutoff on the SIFTER results, which discovers the greatest distance between two posterior probabilities and only outputs the GO terms with a higher posterior probability than this cutoff.

Manual functional annotation in genome projects

As I have shown, at the moment no tool is able to predict the full function of all kinds of genes from any organism with 100% accuracy. To avoid wrong annotations in public databases, which can be propagated through public databases, a manual curation of the automatic annotation is necessary. However this process is very time-consuming because the different analysis results from different tools have to be compared and combined and results are often in different formats. Furthermore each tool has its own cutoffs for trustworthy results and these should also be considered.

To simplify the manual annotation process we have implemented AFAWE, which enables an easy comparison between results from different functional prediction tools by highlighting trustworthy results from each analysis and displaying the results in a way that facilitates the comparison between the results. All analyses in AFAWE are run by web services to ensure the interoperability and the easy extensibility of the system. Currently the newly developed SIFTER workflow together with Blast searches against different databases and tools for protein domain predictions are integrated in AFAWE. After logging in each user is able to add a detailed manual annotation to each protein. To distribute manual annotations to public databases and to build a connection between public databases and the AFAWE system, several web services for data retrieval from the AFAWE database and for starting analyses in AFAWE have been implemented. Furthermore the public database MIPSPlantsDB [Spannagl et al. 2007] has integrated AFAWE protein IDs from the international Medicago genome annotation project IMGAG and from the international tomato genome annotation project ITAG as cross reference, to give the user of MIPSPlantsDB the possibility to go directly from the protein report in MIPSPlantsDB to the corresponding analysis results in AFAWE and add his or her own manual functional annotation. This manual annotation will then be displayed in the protein report of MIPSPlantsDB and will give other users additional information about the function of the protein.

However, although using web services instead of local tools has several advantages like interoperability, scalability, changeability and easy extensibility, they also bring many problems. One problem is the change of the output format, which is returned by the web service. Each time the output format has changed, parsers have to be rewritten and although adapting the parser to a new format is not very time consuming, this requires a full time support, because often the change of the output format is not discovered directly. Mailing lists supported by the provider of the web service could help in this case by announcing the change of the format. Unfortunately these are often not provided or this information is not announced.

Another issue, which causes problems, are dead web services. Especially in science people often change their working environment or the funding for a projects is over. Tools or web services are not supported anymore and this results in many non-functional web services. One way to handle that is by using alternative web services. However, often there is no alternative web service available or the input and output formats of the alternative web service are different, which makes a direct integration of the web service impossible. Another way to avoid dead web services are initiatives like the OMII-UK project in the United Kingdom[26], which supports software if the project runs out of money. Unfortunately OMII-UK supports only UK projects, so corresponding initiatives in others countries are necessary.

Furthermore there are many data retrieval web services currently available, but only few analysis web services. One reason for that are the missing resources at some institutes to handle many different requests and the missing credit for providing a web service. However compute power is becoming cheaper and projects like MyExperiment enable the distribution of web services workflows and the building of new communities and shared workflows. Some projects [Fisher et al. 2007] have already shown that it is possible to get much credit for a community based web service workflow.

Newly discovered protein families

In the potato, Sorghum and Medicago genomes I was able to detect genes, which were not known in plants yet and are therefore potential candidates for further experiments.

[26] *http://www.omii.ac.uk/*

One gene family was only found in potato genomes and so seems to be a potato specific gene family. However, because of missing expression data I am not sure if genes of this family are really expressed. The expression of the genes should be checked before further experiments are done. Nevertheless it is interesting that all genes are located in a hot spot for pathogen resistance and therefore may play a role in this area. To confirm this, phenotype analyses would be necessary. It is also possible that this gene family is a new form of transposon.

Another gene family was first found in Medicago and later in EST data of other plants like citrus, spruce, pine, fir and ferns. The EST data from other plants is an indicator that this gene is likely to be also expressed in Medicago. Genes of this family seem to belong to the family of transferrins, which are well know in animals, insects and some green algae, but no member is known in higher plants yet. Transferrins transfer iron in fluids like human blood. The phylogenetic tree of the transferrin family reflects the evolutionary history from primitive old organisms (e.g. algae, cyanobacteria or ferns) to higher evolved, younger organisms like Angiosperms and insects. Therefore I assume that transferrins are a very old gene family, which is lost in some organisms. This raises the question why only some organisms have transferrins and others do not. One explanation would be a selective advantage. All plant transferrins show a high similarity to transferrin-like genes from algae and insects, which seem to play a role in the innate immune response against bacteria and fungi [Valles et al. 2005] [Thompson et al. 2003]. Iron deprivation, as a result of iron binding proteins like transferrins, prevents the formation of a bacterial biofilm and makes bacteria susceptible to innate immune defense or antibiotics [Ong et al. 2006]. In some insects it is shown that transferrin-like genes are up-regulated during infection [Valles et al. 2005] [Thompson et al. 2003] and in Drosophila they have been shown to be primarily dependent on the Toll-pathway and represent an important iron-withholding strategy [Boutros et al. 2002]. To prove this assumption, further investigations in form of experiments are needed. However, no hints could be found why transferrins are present only in this set of plants and what these plants have in common. A broader view of all plants which include transferrin will be possible if more genomes will become available. However, if it could be shown that transferrins in higher plants are involved in the innate immune response against bacteria or fungi, these genes could be interesting candidates for improving agronomic traits or for the development of resistant varieties.

Also in the Sorghum genome three interesting candidate genes for further experiments could be identified. These genes seem to be a putative horizontal gene transfer from bacteria or come from mitochondrium or chloroplast. Also a calcium binding protein, a putative F-Box protein and an unknown protein could be identified, which are also found in rice, *Vitis vinifera*, populus and *Picea sitchensis*. No hint could be found of the function of these genes and they are also unknown in other plants.

Limitations of the automatic function prediction

I have shown that by using an automatic phylogenomic pipeline it is possible to predict the functions of a protein very precisely. I was able to annotate approximately 55% of the Sorghum genome, 20% of the Medicago genome and 19% of the first sequenced part of the tomato genome. By combining this result with the result from InterProScan in combination with InterPro2GO I was further able to increase the number of annotated genes in the Medicago genome to 33% and in the tomato gene to 35%.

However, still for many genes no GO term could be annotated. One reason for that are errors in the assembly and the gene prediction. Especially in the on-going tomato genome project there are hints to a poor gene prediction, because approximately 30% of all genes show an overlap smaller than 60% with related genes. Because the overlap between tomato sequence and Arabidopsis sequence is quite good, but the overlap between Arabidopsis sequence and tomato sequence is in 30% of the cases below 60%, I assume that many genes are too short or split into two or more genes. In comparison to the SIFTER pipeline, which uses an overlap cutoff of 60% to get candidate homologous genes and therefore was not able to functionally annotate many genes with wrong structural annotations, InterProScan was able to functionally annotate also some of these genes, because tools integrated in InterProScan search for protein domains and functional motifs, but do not consider the overall gene structure. Furthermore some InterProScan tools (e.g. a search with Hidden Markov Models) are more sensitive than a Blast search and are able to additionally recognize incomplete protein domains of genes whose structural annotations are incorrect. However, a functional annotation for structurally mis-annotated genes would not make sense, because if genes are too short, then the functional annotation also becomes incomplete or erroneous. Therefore the structural annotation should be checked before the functional annotation is done. If there are hints to a wrong structural annotation the assigned functional annotation should give a notice of that.

Another reason for the low number of annotated genes are missing GO terms annotated to genes. Still for many genes no GO terms are available or the GO annotation is incomplete [Kourmpetis et al. 2007]. Many genes have a human readable description assigned to describe their function, but no GO term is annotated. Maybe text-mining algorithms can help in the future to translate this description line into GO annotations and therefore enlarge the number of GO annotated genes.

However, the functional prediction becomes hard if homologs of the gene of interest are not functionally characterized yet or no homologous sequence is present in public sequence databases. Also in case of fast evolving genes, like disease resistance genes in plants (see chapter IV), the determination of the function is complicated, because although all genes of this family show a high sequence similarity (~80%) they are often directed against different pathogens and by the transfer of function only the information that this is a disease resistance gene can be retrieved, but no information about the pathogen, against which the gene is directed, can be obtained. In this case transferring the function from one gene to another is not possible and other methods should be used instead like e.g. looking for co-expressed genes, comparing the shared synteny and the structure and search for functionally known motifs (e.g. active sites). But in many cases also these results will give no hint to the function of the protein and further investigation in the laboratory is necessary. However, in genome projects running all kinds of analyses would be too time consuming and the transfer of function gives a first idea of the content of the genome.

Another bottleneck is the description of the function. For this case many ontologies are available, which avoid synonyms and make a description of the function of a gene machine readable. Probably the most common function vocabulary is Gene Ontology, which is also used in this thesis. However, although Gene Ontology is machine readable and enables the fast comparison between function terms, it also has limitations. One of them is that Gene Ontology is incomplete for many organisms and several terms are missing (e.g. functions of plant transcription factors). In this case other vocabularies like MapMan [Thimm et al. 2004] could help to define the function of a protein more precisely. Another problem is that Gene Ontology sometimes provides a name of a term (e.g. muscle alpha-actinin binding), which is true for one organism group (animals), but not for another (plants also have actin, but no muscles). So by transferring a GO term from one gene to another it has to be considered if the corresponding term is applicable for this organism. Furthermore Gene Ontology is also work in progress and in each release terms become obsolete and new terms are created. It is also important to take care of that and to use only up-to-date data.

IX. Outlook

In the post-genomic era the determination of the function of genes is the next big challenge. In this thesis I have developed tools and workflows for the automatic and manual annotation of proteins and these perform very well on the tested datasets. However these tools also have limitations and can be improved and extended.

One extension could be the integration of expression data and structure data in SIFTER-X. But expression data should be only used for the prediction of the biological process and not for the molecular function, because proteins can be differently expressed, but have the same molecular function. Structure data could be used on the other hand to find out, if changes in the amino acid sequence are important for the function of the protein or not. If these critical residues have changed between two nodes the function should not be transferred between nodes.

Another extension could be the integration of a profile search in the phylogenomic pipeline (as mentioned in the discussion part) to get additional homologous genes for the phylogenetic tree. Also instead of the manually built RefSeq database, which includes only fully sequenced genes, a more comprehensive database like UniProt should be used. In this case the additional functional information from not fully sequenced organisms can be obtained. However this will result in a decrease in speed and in a decreased accuracy of the phylogenetic tree because of incomplete genomes.

With regard to manual annotation, I have provided with AFAWE the first prototype to facilitate a fast comparison between analysis results and to add a manual annotation. Although AFAWE is very intuitive, has already some users and some manual annotations have been made, it has to be further promoted to inform potential new users about its existence, show them how to use it and make it popular. To provide the user of AFAWE with a broad selection of analyses for comparison, other tools like the display of co-expressed genes or information about the structure of the protein should be integrated. Another idea is also to integrate the promoter sequence of the gene of interest in AFAWE and provide tools to automatically find motifs, which are conserved between promoters of homologous genes. Additionally the comparison of the analysis results should be simplified by displaying summaries of Ontology terms of each analysis individually and a combination of all of them on a separate summary page. To overcome problems with dead or temporarily unavailable web services, alternative web services could be called.

X. Appendix

1) 100 manual inspected Medicago gene predictions made by SIFTER

IMGAG (1.0) Identifier	***Predicted GO term by SIFTER***	***True Prediction?***	***Annotated description line***	***Is the SIFTER prediction more specific as the annotated description line?***	***Comments***
IMGA| AC139354_35.1	0005515 (protein binding)	YES	ChaC-like protein	-	
IMGA| AC150703_2.1	0004623 (phospholipase A2 activity)	YES	Phospholipase A2	-	
IMGA| CR931741_6.2	0045551 (cinnamyl-alcohol dehydrogenase activity)	YES	Alcohol dehydrogenase superfamily, zinc-containing	YES	
IMGA| AC173834_25.1	0004872 (receptor activity)	YES	Bacteriophytochrome	-	
IMGA| AC143341_25.2	0003677 (DNA binding)	YES	Homeodomain-like	YES	
IMGA| AC146575_6.2	0005525 (GTP binding)	YES	Ras small GTPase, Ras type; Small GTP-binding protein domain	-	
IMGA| AC151709_22.2	0004046 (aminoacylase activity)	YES	Peptidase dimerisation	YES	
IMGA| AC144389_35.2	0004129 (cytochrome-c oxidase activity)	NO	Cytochrome b5	-	
IMGA| AC126778_7.2	0009044 (xylan 1,4-beta-xylosidase activity)	YES	Glycoside hydrolase, family 3, N-terminal; Glycoside hydrolase, family 3, C-terminal	YES	
IMGA| AC174370_24.1	0004055 (argininosuccinate synthase activity)	YES	Argininosuccinate synthase	-	
IMGA| AC148481_30.2	0047209 (coniferyl-alcohol glucosyltransferase activity)	YES	UDP-glucoronosyl and UDP-glucosyl transferase family protein	-	

IMGAG (1.0) Identifier	***Predicted GO term by SIFTER***	***True Prediction?***	***Annotated description line***	***Is the SIFTER prediction more specific as the annotated description line?***	***Comments***
IMGA| AC136507_24.2	0008565 (protein transporter activity)	YES	Longin-like	NO	
IMGA| AC149803_8.2	0045431 (flavonol synthase activity)	NO	2OG-Fe(II) oxygenase	-	
IMGA| CR954193_19.2	0005509 (calcium ion binding)	YES	Calcium-binding EF-hand	-	
IMGA| AC141111_15.2	0008453 (alanine-glyoxylate transaminase activity)	YES	Aminotransferase, class V	YES	
IMGA| AC158546_1.1	0003700 (transcription factor activity)	YES	Zinc finger, Dof-type	YES	
IMGA| AC124957_30.2	0005102 (receptor binding)	YES	Quinonprotein alcohol dehydrogenase-like	YES	Description is wrong
IMGA| AC152177_43.1	0015297 (antiporter activity)	YES	Multi antimicrobial extrusion protein MatE	-	
IMGA| AC140026_13.2	0005554 (unknown)	YES	Protein of unknown function DUF239	-	Unknown protein
IMGA| AC144766_8.2	0003700 (transcription factor activity)	YES	Zinc finger, CCCH-type	YES	
IMGA| CR936327_2.2	0004218 (cathepsin S activity)	?	Peptidase C1A, papain; Peptidase M14, carboxypeptidase A	-	Not sure if true or wrong
IMGA| AC174346_30.1	0047251 (thiohydroximate beta-D-glucosyltransferase activity)	YES	UDP-glucoronosyl and UDP-glucosyl transferase family protein	-	
IMGA| CT863712_13.1	0003700 (transcription factor activity)	YES	Transcription factor, MADS-box	-	
IMGA| AC165276_13.1	0000036 (acyl carrier	YES	Acyl carrier protein (ACP)	-	

IMGAG (1.0) Identifier	***Predicted GO term by SIFTER***	***True Prediction?***	***Annotated description line***	***Is the SIFTER prediction more specific as the annotated description line?***	***Comments***
	activity)				
IMGA| AC157646_18.1	0005427 (proton-dependent oligopeptide secondary active transmembrane transporter activity)	YES	TGF-beta receptor, type I/II extracellular region	-	
IMGA| AC149578_26.2	0045735 (nutrient reservoir activity)	YES	Cupin region	YES	
IMGA| AC133862_11.2	0051082 (unfolded protein binding)	YES	Heat shock protein DnaJ	NO	
IMGA| AC169177_28.1	0005515 (protein binding)	YES	Leucine-rich repeat	-	
IMGA| CT573052_26.2	0015095 (magnesium ion transmembrane transporter activity)	YES	Mg2+ transporter protein, CorA-like	-	
IMGA| AC127429_17.2	0045735 (nutrient reservoir activity)	YES	BURP	-	
IMGA| AC174289_17.1	0030528 (transcription regulator activity)	YES	Protein of unknown function DUF581	YES	Description wrong
IMGA| CT573028_11.2	0004805 (trehalose-phosphatase activity)	YES	Trehalose-phosphatase	-	
IMGA| AC155896_11.2	0046872 (metal ion binding)	YES	WD40-like	-	
IMGA| CT027660_11.1	0003841 (1-acylglycerol-3-phosphate O-acyltransferase activity)	YES	1-acyl-sn-glycerol-3-phosphate acyltransferase	-	

IMGAG (1.0) Identifier	***Predicted GO term by SIFTER***	***True Prediction?***	***Annotated description line***	***Is the SIFTER prediction more specific as the annotated description line?***	***Comments***
IMGA|AC173964_29.1	0005529 (sugar binding)	YES	Protein kinase; Curculin-like (mannose-binding) lectin	NO	
IMGA|AC124214_16.2	0030060 (L-malate dehydrogenase activity)	YES	Lactate/malate dehydrogenase	-	
IMGA|AC119413_42.2	0003723 (RNA binding)	YES	RNA-binding region RNP-1 (RNA recognition motif)	-	
IMGA|AC153354_6.1	0005198 (structural molecule activity)	YES	Initiation factor eIF-4 gamma, middle; Initiation factor eIF-4 gamma, MA3	NO	
IMGA|CR954198_13.2	0005515 (protein binding)	YES	FAR1; Zinc finger, SWIM-type; Cupin, RmlC-type	NO	-
IMGA|AC126782_49.2	0016887 (ATPase activity)	YES	AAA ATPase, central region; SMAD/FHA	-	-
IMGA|AC151621_30.1	0003730 (mRNA 3'-UTR binding)	YES	RNA-binding region RNP-1 (RNA recognition motif)	-	-
IMGA|CT573052_7.2	0003729 (mRNA binding)	YES	Pentatricopeptide repeat	YES	-
IMGA|AC133780_12.1	0004674 (protein serine/threonine kinase activity)	YES	Protein kinase	YES	-
IMGA|AC134049_54.2	0005524 (ATP binding)	YES	Disease resistance protein	NO	-
IMGA|AC161033_18.2	0051087 (chaperone binding)	YES	Heat shock protein DnaJ	NO	-
IMGA|AC122164_6.2	0005524 (ATP binding)	YES	Protein kinase	NO	-
IMGA|AC148487_4.2	0047652 (allantoate deiminase activity)	YES	Peptidase M20	YES	-
IMGA|AC146557_11.1	0004815 (aspartate-tRNA ligase activity)	YES	GAD; Aminoacyl-transfer RNA synthetase, class II	-	-

IMGAG (1.0) Identifier	***Predicted GO term by SIFTER***	***True Prediction?***	***Annotated description line***	***Is the SIFTER prediction more specific as the annotated description line?***	***Comments***
IMGA|AC147499_13.2	0005515 (protein binding)	?	Thaumatin, pathogenesis-related	NO	Not sure
IMGA|AC174315_5.1	0005516 (calmodulin binding)	YES	IQ calmodulin-binding region	-	-
IMGA|AC124952_6.2	0019153 (protein-disulfide reductase (glutathione) activity)	YES	Thioredoxin domain 2	YES	-
IMGA|AC129090_32.2	0005554 (unknown function)	-	HMG-I and HMG-Y, DNA-binding	YES	Description is wrong
IMGA|AC139290_19.2	0004687 (myosin light chain kinase activity)	YES	Protein kinase	YES	-
IMGA|AC166038_4.1	0009000 (selenocysteine lyase activity)	YES	MOSC; MOSC, N-terminal beta barrel	YES	1 GO term is missing
IMGA|AC140032_7.1	0016165 (lipoxygenase activity)	YES	Lipoxygenase	-	-
IMGA|AC155880_17.2	0005515 (protein binding)	YES	HAD-superfamily subfamily IB hydrolase, hypothetical	NO	-
IMGA|CT030192_7.1	0004842 (ubiquitin-protein ligase activity)	YES	Cyclin-like F-box; F-box protein interaction domain; Galactose oxidase, central	YES	-
IMGA|AC149038_17.2	0005554 (unknown function)	-	Protein of unknown function	-	Unknown protein
IMGA|AC144539_41.2	0003700 (transcription factor activity)	YES	GRAS transcription factor	NO	-
IMGA|AC146720_15.2	0005554 (unknown function)	-	Leucine-rich repeat; Leucine-rich repeat, cysteine-containing type	NO	-
IMGA|AC167958_2.1	0005524 (ATP binding)	YES	EMB1135; ATP binding , putative	-	-

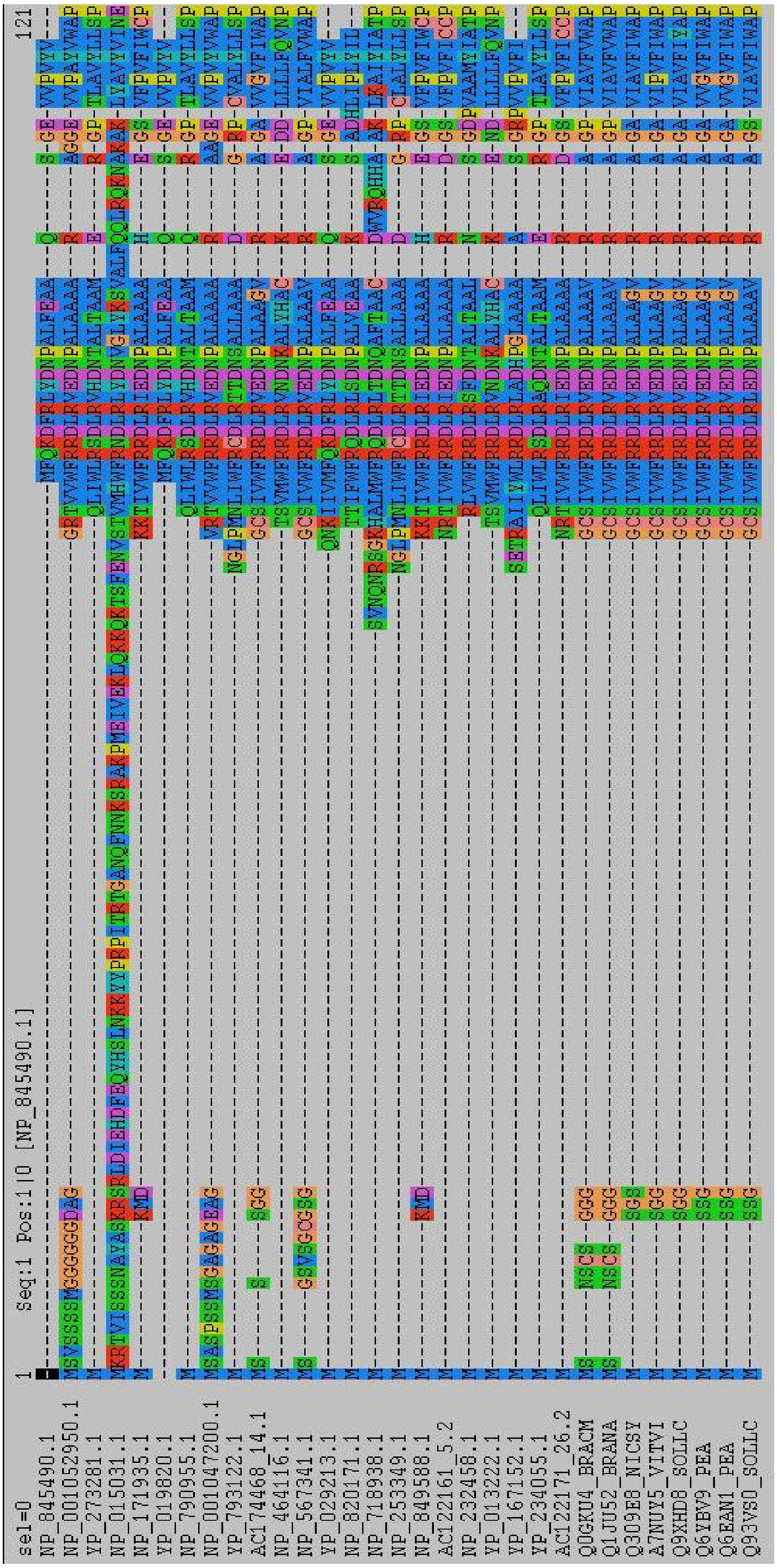
IMGAG (1.0) Identifier	***Predicted GO term by SIFTER***	***True Prediction?***	***Annotated description line***	***Is the SIFTER prediction more specific as the annotated description line?***	***Comments***
IMGA|AC144502_17.2	0008270 (zinc ion binding)	YES	Zinc finger, C2H2-type	-	-
IMGA|AC152552_56.1	0005524 (ATP binding)	YES	Protein kinase	-	1 GO term is missing
IMGA|AC148097_6.2	0016301 (kinase activity)	YES	Protein kinase	-	-
IMGA|AC169089_8.1	0030515 (snoRNA binding)	YES	Fibrillarin	NO	-
IMGA|AC141113_49.2	0003700 (transcription factor activity)	YES	DNA-binding WRKY	-	-
IMGA|AC122730_40.2	0004867 (serine-type endopeptidase inhibitor activity)	YES	Kunitz inhibitor ST1-like	-	1 GO term is missing
IMGA|AC157348_18.1	0003924 (GTPase activity)	YES	Ras GTPase; Calcium-binding EF-hand	-	-
IMGA|CT010481_7.2	0004367 (glycerol-3-phosphate dehydrogenase (NAD+) activity)	YES	NAD-dependent glycerol-3-phosphate dehydrogenase, C-terminal	-	-
IMGA|AC152552_10.1	0030508 (thiol-disulfide exchange intermediate activity)	YES	Thioredoxin domain 2; Thioredoxin fold	YES	1 GO term is missing
IMGA|AC126012_2.2	0005515 (protein binding)	YES	Peptidase S59, nucleoporin	NO	-
IMGA|AC157983_25.2	0016855 (racemase and epimerase activity, acting on amino acids and derivatives)	YES	Asp/Glu racemase	-	-
IMGA|AC149131_5.2	0001758 (retinal dehydrogenase activity)	NO	Short-chain dehydrogenase/reductase SDR	-	Description is unspecific
IMGA|CU013514_6.1	0003700 (transcription factor activity)	YES	Zinc finger, CCHC-type; Homeodomain-related	YES	-

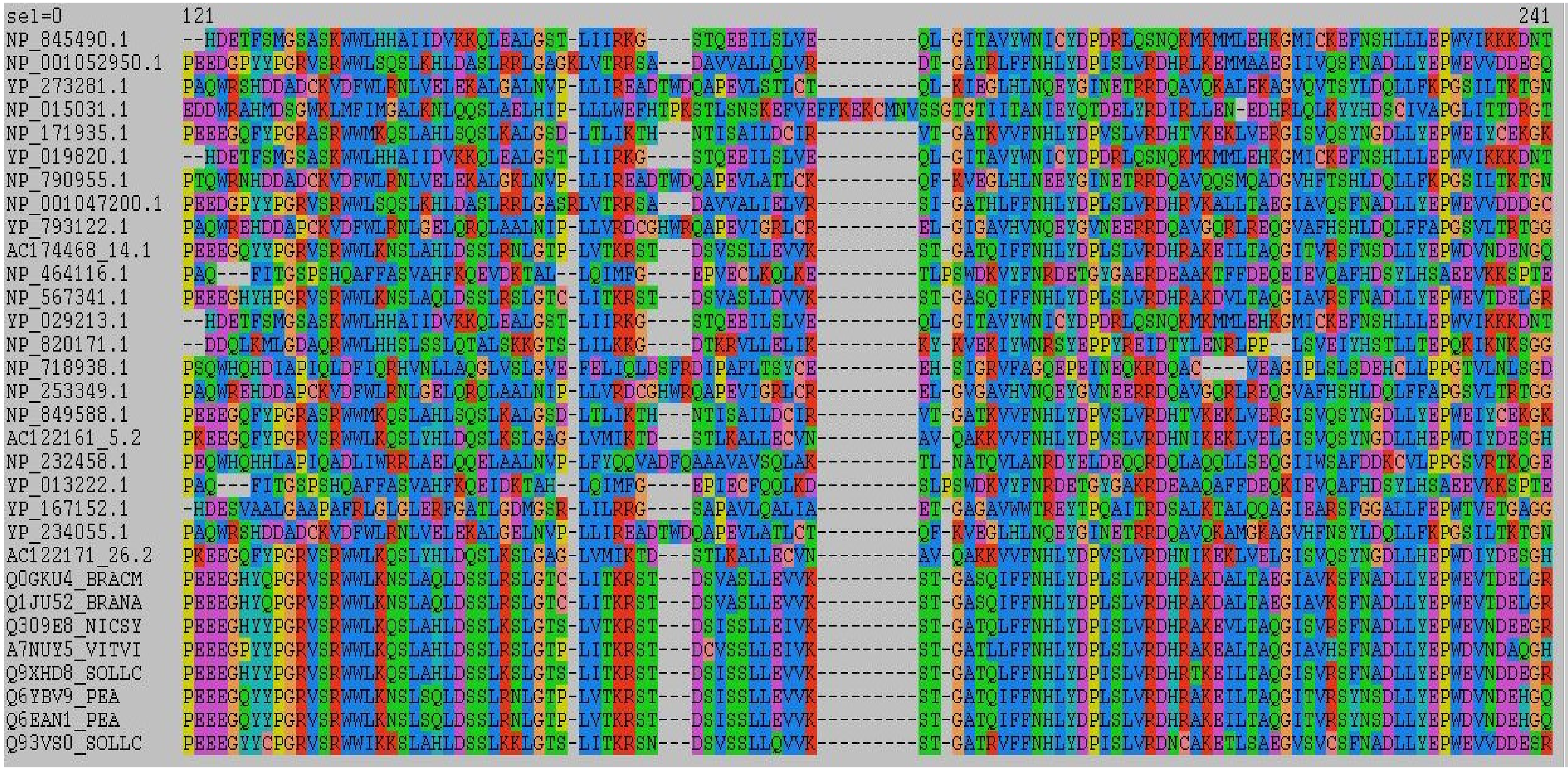
IMGAG (1.0) Identifier	***Predicted GO term by SIFTER***	***True Prediction?***	***Annotated description line***	***Is the SIFTER prediction more specific as the annotated description line?***	***Comments***
IMGA|AC135797_6.2	0003743 (translation initiation factor activity)	YES	Proteasome component region PCI	YES	Description is wrong
IMGA|AC146748_15.2	0008134 (transcription factor binding)	YES	TGS; Small GTP-binding protein domain	YES	1 GO term is missing
IMGA|CU012059_15.1	0003735 (structural constituent of ribosome)	YES	Ribosomal protein S12, bacterial and chloroplast form	-	-
IMGA|AC126783_10.2	0004497 (monooxygenase activity)	YES	FAD-dependent pyridine nucleotide-disulphide oxidoreductase	YES	Description is wrong
IMGA|AC146586_38.2	0005524 (ATP binding)	YES	Protein kinase	-	1 GO term is missing
IMGA|AC151526_7.2	0016787 (hydrolase activity)	YES	Histidine acid phosphatase; HAD-superfamily hydrolase subfamily IA, variant 3	YES	Description is wrong
IMGA|AC163383_6.1	0051082 (unfolded protein binding)	YES	Heat shock protein DnaJ, N-terminal; Homeodomain-related	NO	-
IMGA|AC146705_13.2	0005516 (calmodulin binding)	?	Auxin responsive SAUR protein	-	-
IMGA|CT954231_4.2	0003824 (catalytic activity)	YES	Metal-dependent phosphohydrolase, HD region	NO	-
IMGA|AC143341_4.2	0005554 (unknown function)	-	hypothetical protein	-	Unknown function
IMGA|AC148918_38.2	0004523 (ribonuclease H activity)	YES	Polynucleotidyl transferase, Ribonuclease H fold	-	-
IMGA|AC174362_13.1	0003735 (structural constituent of ribosome)	YES	Ribosomal protein S19/S15	-	-
IMGA|AC166897_16.1	0003824 (catalytic activity)	YES	Protein of unknown function DUF676, hydrolase-like	-	-

IMGAG (1.0) Identifier	***Predicted GO term by SIFTER***	***True Prediction?***	***Annotated description line***	***Is the SIFTER prediction more specific as the annotated description line?***	***Comments***
IMGA|CT010459_6.2	0046592 (polyamine oxidase activity)	YES	NAD-binding site	YES	-
IMGA|AC149471_8.1	0008320 (protein transmembrane transporter activity)	YES	Importin-beta, N-terminal	YES	-
IMGA|AC148236_12.1	0016829 (lyase activity)	YES	Carboxypeptidase regulatory region; Rhamnogalacturonate lyase	NO	-
IMGA|AC166743_6.1	0004004 (ATP-dependent RNA helicase activity)	YES	Helicase, C-terminal; Zinc finger, CCHC-type	YES	-
IMGA|AC129090_53.2	0019825 (oxygen binding)	YES	Globin; Globin-related	-	-
IMGA|AC141112_9.2	0015359 (amino acid transmembrane transporter activity)	YES	Amino acid/polyamine transporter II	-	-
IMGA|AC146307_20.1	0051082 (unfolded protein binding)	YES	Heat shock protein Hsp20	-	1 GO term is missing
IMGA|AC130803_6.1	0004553 (hydrolase activity, hydrolyzing O-glycosyl compounds)	YES	Glycoside hydrolase, family 5	-	-
IMGA|AC155100_12.1	0016168 (chlorophyll binding)	YES	Chlorophyll A-B binding protein	-	-
IMGA|AC165943_7.1	0005515 (protein binding)	YES	Reticulon	NO	-
IMGA|AC174360_8.2	0008173 (RNA methyltransferase activity)	YES	tRNA/rRNA methyltransferase (SpoU)	-	-
IMGA|CR956402_6.1	0005524 (ATP binding)	YES	Disease resistance protein	NO	-

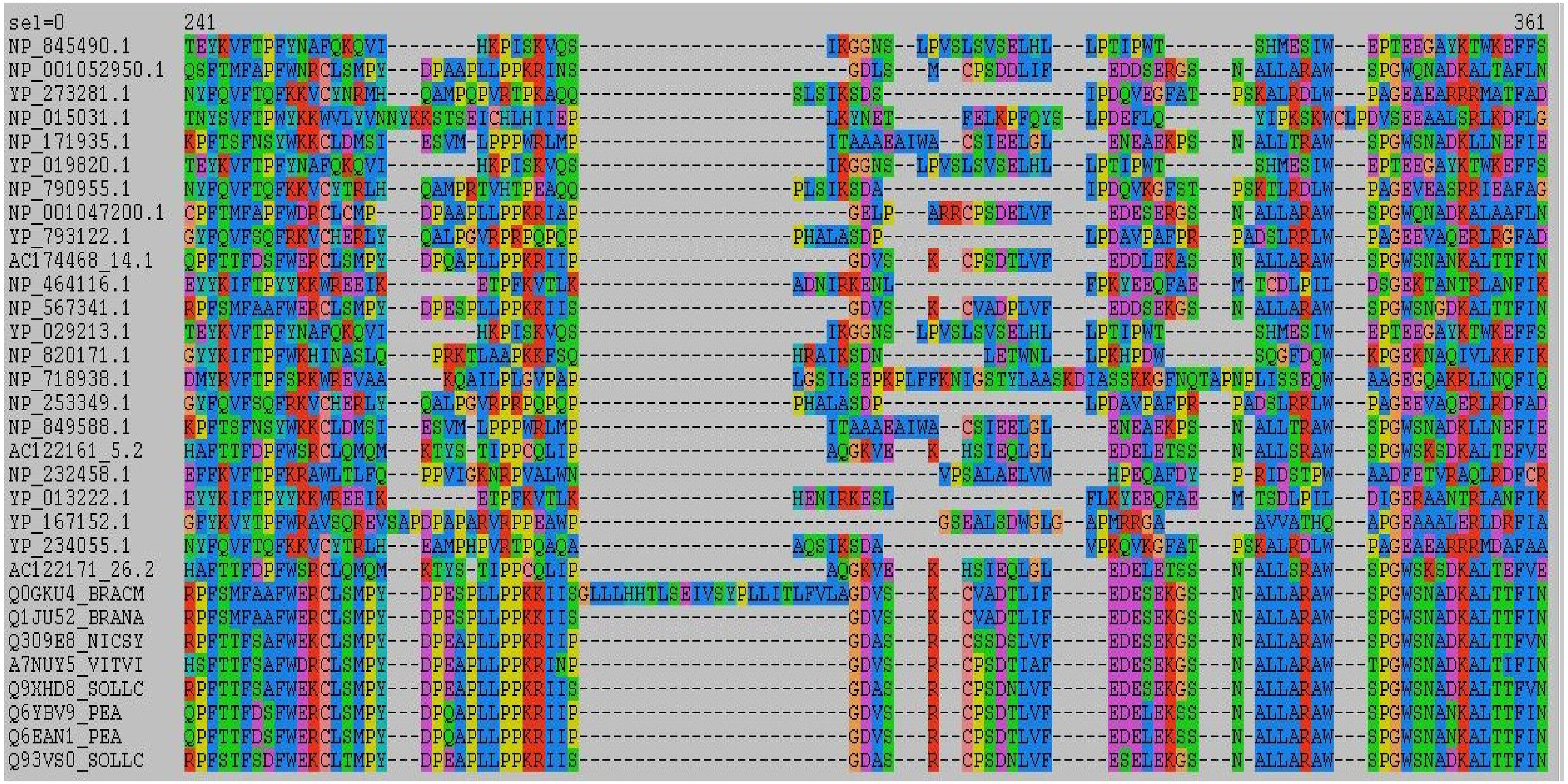
IMGAG (1.0) Identifier	***Predicted GO term by SIFTER***	***True Prediction?***	***Annotated description line***	***Is the SIFTER prediction more specific as the annotated description line?***	***Comments***
IMGA| CR931734_3.2	0005515 (protein binding)	YES	FAR1; Zinc finger, SWIM-type	-	-

2) Alignment of the photolyase/blue-light photoreceptor family

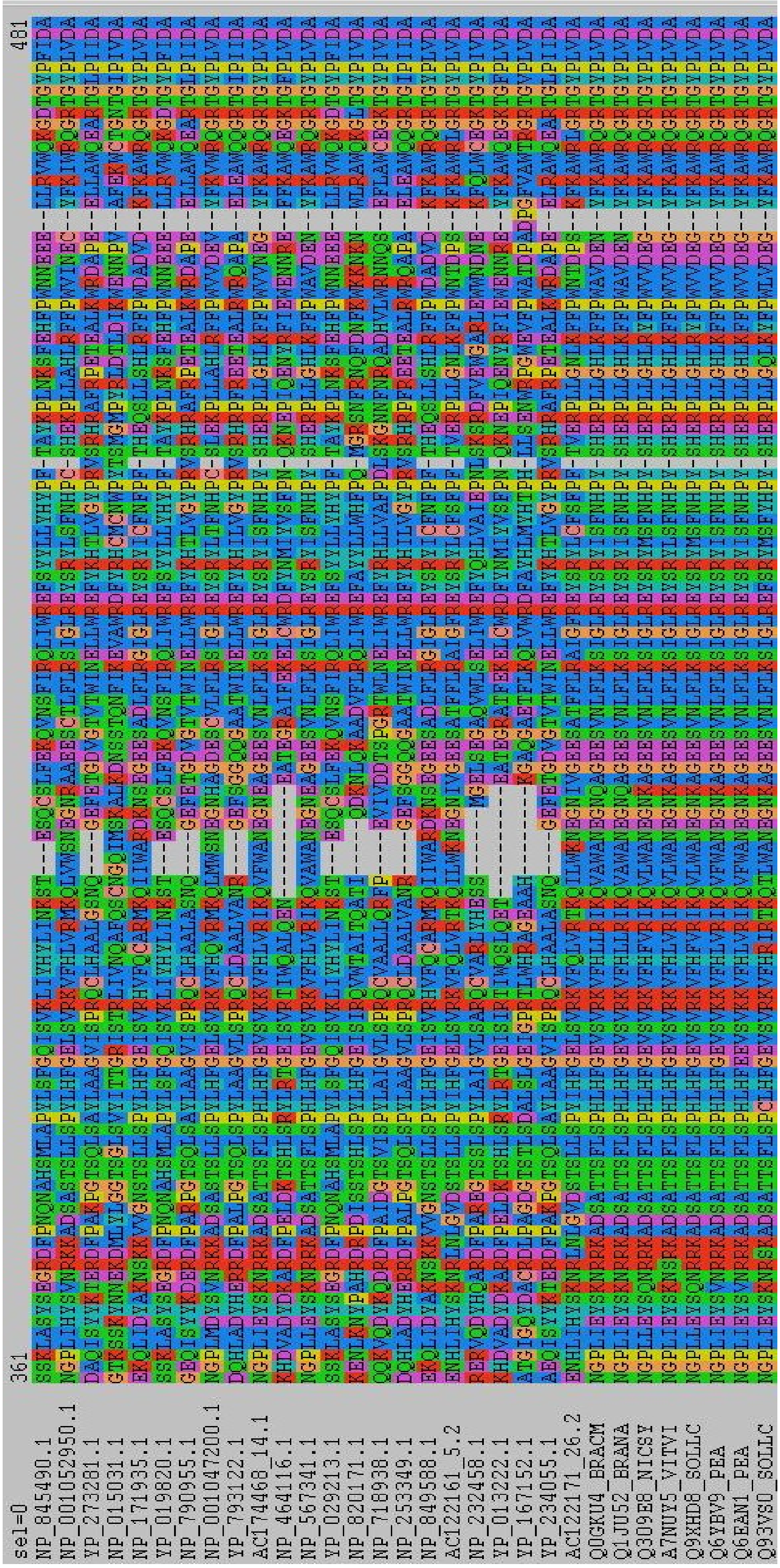


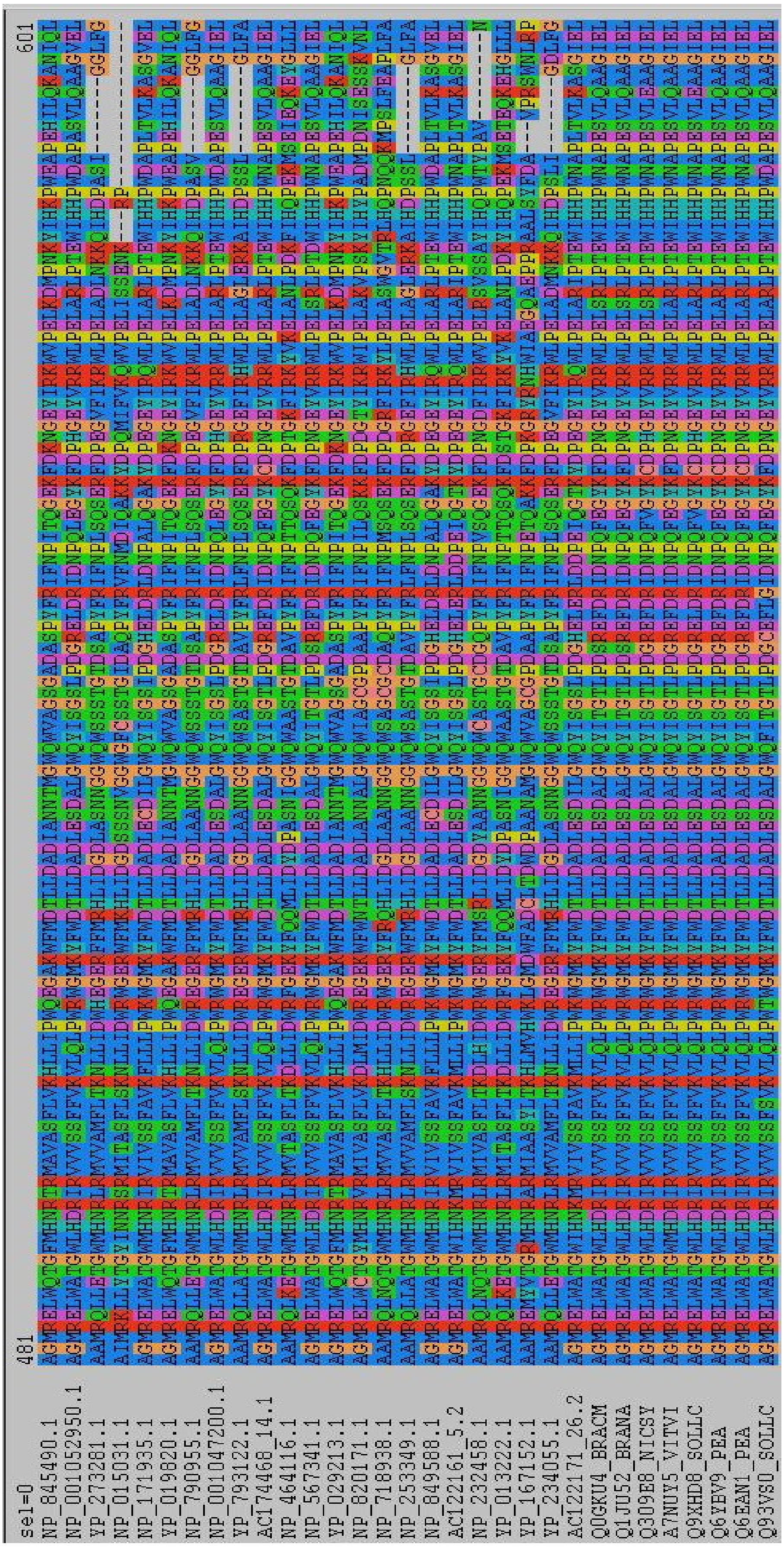


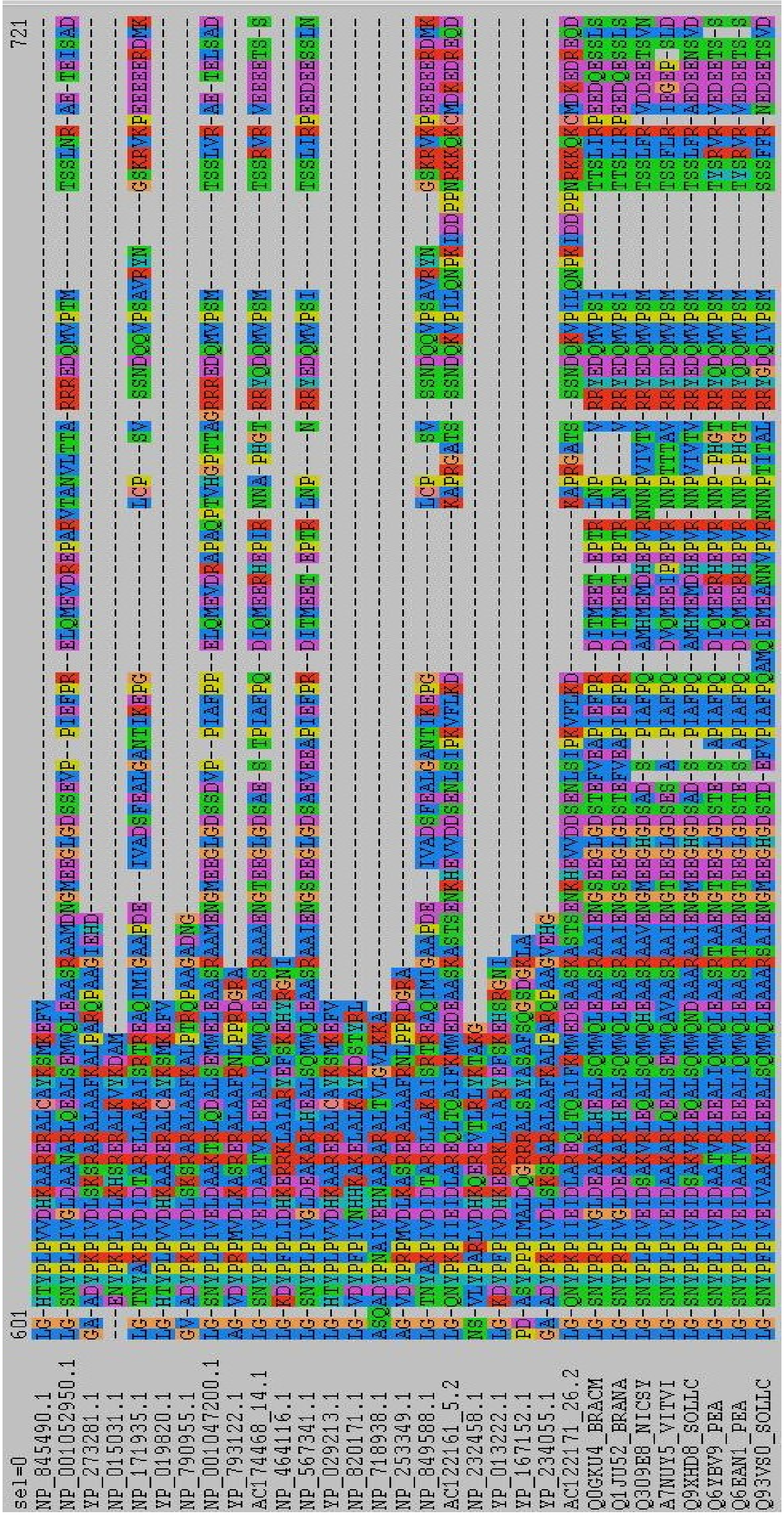
```
sel=0            121                                                                                                                241
NP_845490.1      --HDETFSMGSASKWWLHHAIIDVKKQLEALGST-LIIRKG----STQEEILSLVE---------QL-GITAVYWNICYDPDRLQSNQKMKMMLEHKGMICKEFNSHLLLEPWVIKKKDNT
NP_001052950.1   PEEDGPYYPGRVSRWWLSQSLKHLDASLRRLGAGKLVTRRSA---DAVVALLQLVR---------DT-GATRLFFNHLYDPISLVRDHRLKEMMAAEGIIVQSFNADLLYEPWEVVDDEGQ
YP_273281.1      PAQWRSHDDADCKVDFWLRNLVELEKALGALNVP-LLIREADTWDQAPEVLSTLCT---------QL-KIEGLHLNQEYGINETRRDQAVQKALEKAGVQVTSYLDQLLFKPGSILTKTGN
NP_015031.1      EDDWRAHMDSGWKLMFIMGALKNLQQSLAELHIP-LLLWEFHTPKSTLSNSKEFVEFFKEKCMNVSSGTGTIITANIEYQTDELYRDIRLLEN-EDHRLQLKYYHDSCIVAPGLITTDRGT
NP_171935.1      PEEEGQFYPGRASRWWMKQSLAHLSQSLKALGSD-LTLIKTH---NTISAILDCIR---------VT-GATKVVFNHLYDPVSLVRDHTVKEKLVERGISVQSYNGDLLYEPWEIYCEKGK
YP_019820.1      --HDETFSMGSASKWWLHHAIIDVKKQLEALGST-LIIRKG----STQEEILSLVE---------QL-GITAVYWNICYDPDRLQSNQKMKMMLEHKGMICKEFNSHLLLEPWVIKKKDNT
NP_790955.1      PTQWRNHDDADCKVDFWLRNLVELEKALGKLNVP-LLIREADTWDQAPEVLATLCK---------QF-KVEGLHLNEEYGINETRRDQAVQQSMQADGVHFTSHLDQLLFKPGSILTKTGN
NP_001047200.1   PEEDGPYYPGRVSRWWLSQSLKHLDASLRRLGASRLVTRRSA---DAVVALIELVR---------SI-GATHLFFNHLYDPLSLVRDHRVKALLTAEGIAVQSFNADLLYEPWEVVDDDGC
YP_793122.1      PAQWREHDDAPCKVDFWLRNLGELQRQLAALNIP-LLVRDCGHWRQAPEVIGRLCR---------EL-GIGAVHVNQEYGVNEERRDQAVGQRLREQGVAFHSHLDQLFFAPGSVLTRTGG
AC174468_14.1    PEEEGQYYPGRVSRWWLKNSLAHLDSSLRNLGTP-LVTKRST---DSVSSLLEVVK---------ST-GATQIFFNHLYDPLSLVRDHRAKEILTAQGITVRSFNSDLLYEPWDVNDENGQ
NP_464116.1      PAQ---FITGSPSHQAFFASVAHFKQEVDKTAL--LQIMFG----EPVECLKQLKE---------TLPSWDKVYFNRDETGYGAERDEAAKTFFDEQEIEVQAFHDSYLHSAEEVKKSPTE
NP_567341.1      PEEEGHYHPGRVSRWWLKNSLAQLDSSLRSLGTC-LITKRST---DSVASLLDVVK---------ST-GASQIFFNHLYDPLSLVRDHRAKDVLTAQGIAVRSFNADLLYEPWEVTDELGR
YP_029213.1      --HDETFSMGSASKWWLHHAIIDVKKQLEALGST-LIIRKG----STQEEILSLVE---------QL-GITAVYWNICYDPDRLQSNQKMKMMLEHKGMICKEFNSHLLLEPWVIKKKDNT
NP_820171.1      --DDQLKMLGDAQRWWLHHSLSSLQTALSKKGTS-LILKKG----DTKRVLLELIK---------KY-KVEKIYWNRSYEPPYREIDTYLENRLPP--LSVEIYHSTLLTEPQKIKNKSGG
NP_718938.1      PSQWHQHDIAPIQLDFIQRHVNLLAQGLVSLGVE-FELIQLDSFRDIPAFLTSYCE---------EH-SIGRVFAGQEPEINEQKRDQAC----VEAGIPLSLSDEHCLLPPGTVLNLSGD
NP_253349.1      PAQWREHDDAPCKVDFWLRNLGELQRQLAALNIP-LLVRDCGHWRQAPEVIGRLCR---------EL-GVGAVHVNQEYGVNEERRDQAVGQRLREQGVAFHSHLDQLFFAPGSVLTRTGG
NP_849588.1      PEEEGQFYPGRASRWWMKQSLAHLSQSLKALGSD-LTLIKTH---NTISAILDCIR---------VT-GATKVVFNHLYDPVSLVRDHTVKEKLVERGISVQSYNGDLLYEPWEIYCEKGK
AC122161_5.2     PKEEGQFYPGRVSRWWLKQSLYHLDQSLKSLGAG-LVMIKTD---STLKALLECVN---------AV-QAKKVVFNHLYDPVSLVRDHNIKEKLVELGISVQSYNGDLLHEPWDIYDESGH
NP_232458.1      PEQWHQHHLAPIQADLIWRRLAELQQELAALNVP-LFYQQVADFQAAAVAVSQLAK---------TL-NATQVLANRDYELDEQQRDQLAQQLLSEQGIIWSAFDDKCVLPPGSVRTKQGE
YP_013222.1      PAQ---FITGSPSHQAFFASVAHFKQEIDKTAH--LQIMFG----EPIECFQQLKD---------SLPSWDKVYFNRDETGYGAKRDEAAQAFFDEQKIEVQAFHDSYLHSAEEVKKSPTE
YP_167152.1      -HDESVAALGAAPAFRLGLGLERFGATLGDMGSR-LILRRG----SAPAVLQALIA---------ET-GAGAVWWTREYTPQAITRDSALKTALQQAGIEARSFGGALLFEPWTVETGAGG
YP_234055.1      PAQWRSHDDADCKVDFWLRNLVELEKALGELNVP-LLIREADTWDQAPEVLATLCT---------QF-KVEGLHLNQEYGINETRRDQAVQKAMGKAGVHFNSYLDQLLFKPGSILTKTGN
AC122171_26.2    PKEEGQFYPGRVSRWWLKQSLYHLDQSLKSLGAG-LVMIKTD---STLKALLECVN---------AV-QAKKVVFNHLYDPVSLVRDHNIKEKLVELGISVQSYNGDLLHEPWDIYDESGH
Q0GKU4_BRACM     PEEEGHYQPGRVSRWWLKNSLAQLDSSLRSLGTC-LITKRST---DSVASLLEVVK---------ST-GASQIFFNHLYDPLSLVRDHRAKDALTAEGIAVKSFNADLLYEPWEVTDELGR
Q1JU52_BRANA     PEEEGHYQPGRVSRWWLKNSLAQLDSSLRSLGTC-LITKRST---DSVASLLEVVK---------ST-GASQIFFNHLYDPLSLVRDHRAKDALTAEGIAVKSFNADLLYEPWEVTDELGR
Q309E8_NICSY     PEEEGHYYPGRVSRWWLKQSLAHLDSSLKSLGTS-LVTKRST---DSISSLLEIVK---------ST-GATQLFFNHLYDPISLVRDHRAKEVLTAQGISVRSFNADLLYEPWEVNDEEGR
A7NUY5_VITVI     PEEEGPYYPGRVSRWWLKQSLAHLDSSLRSLGTP-LITKRST---DCVSSLLEIVK---------ST-GATLLFFNHLYDPLSLVRDHRAKEALTAQGIAVHSFNADLLYEPWDVNDAQGH
Q9XHD8_SOLLC     PEEEGHYYPGRVSRWWLKQSLAHLDSSLKSLGTS-LITKRST---DSISSLLEVVK---------ST-GATQLFFNHLYDPISLVRDHRTKEILTAQGISVRSFNADLLYEPWEVNDDEGR
Q6YBV9_PEA       PEEEGQYYPGRVSRWWLKNSLSQLDSSLRNLGTP-LVTKRST---DSISSLLEVVK---------ST-GATQIFFNHLYDPLSLVRDHRAKEILTAQGITVRSYNSDLLYEPWDVNDEHGQ
Q6EAN1_PEA       PEEEGQYYPGRVSRWWLKNSLSQLDSSLRNLGTP-LVTKRST---DSISSLLEVVK---------ST-GATQIFFNHLYDPLSLVRDHRAKEILTAQGITVRSYNSDLLYEPWDVNDEHGQ
Q93VS0_SOLLC     PEEEGYYCPGRVSRWWIKKSLAHLDSSLKKLGTS-LITKRSN---DSVSSLLQVVK---------ST-GATRVFFNHLYDPISLVRDNCAKETLSAEGVSVCSFNADLLYEPWEVVDDESR
```

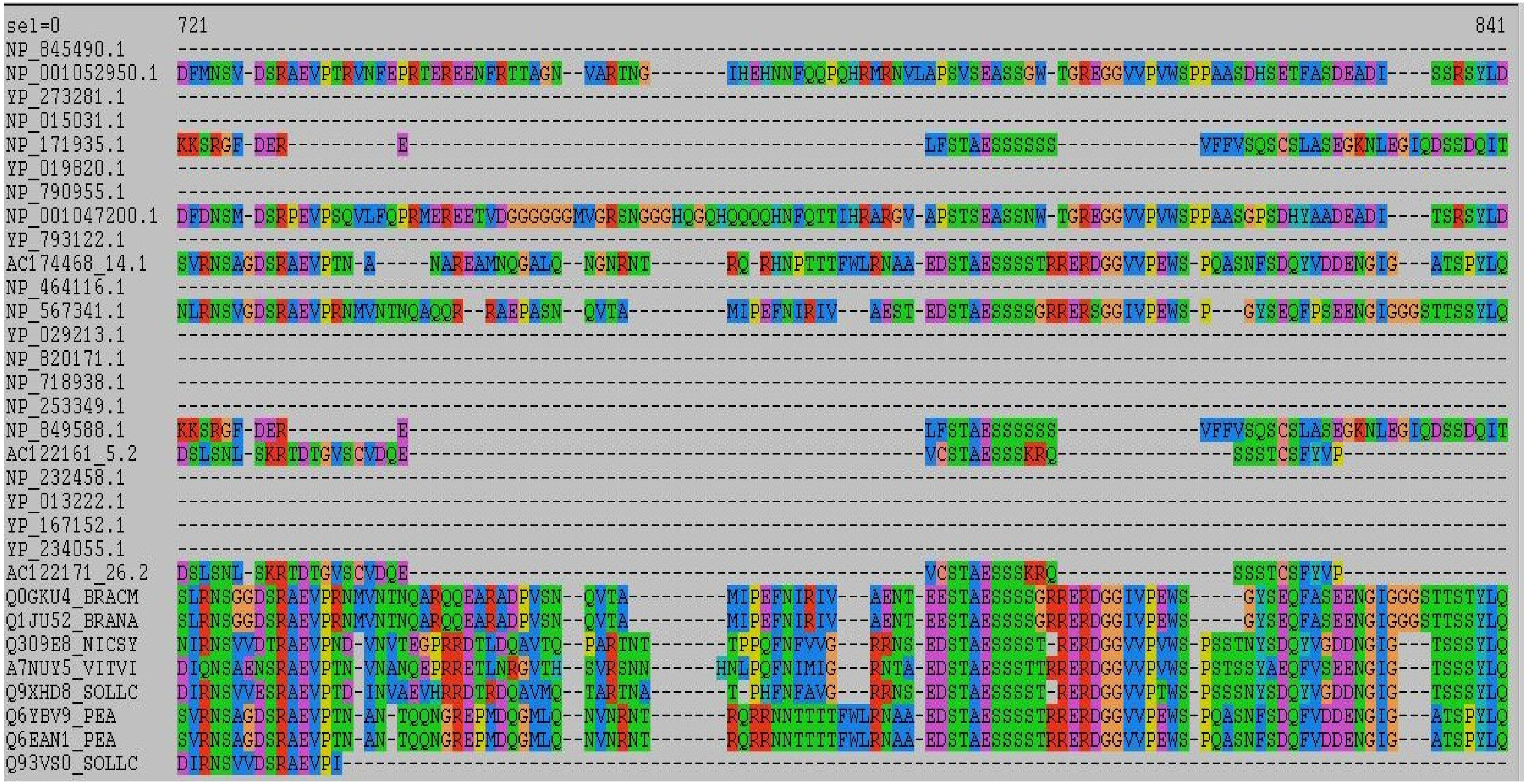


```
sel=0            241                                                                                                                   361
NP_845490.1      TEYKVFTPFYNAFQKQVI--------HKPISKVQS----------------------IKGGNS--LPVSLSVSELHL---LPTIPWT--------SHMESIW---EPTEEGAYKTWKEFFS
NP_001052950.1   QSFTMFAPFWNRCLSMPY---DPAAPLLPPKRINS------------------------GDLS---M--CPSDDLIF-----EDDSERGS---N-ALLARAW---SPGWQNADKALTAFLN
YP_273281.1      NYFQVFTQFKKVCYNRMH---QAMPQPVRTPKAQQ-------------------SLSIKSDS------------------IPDQVEGFAT---PSKALRDLW---PAGEAEARRRMATFAD
NP_015031.1      TNYSVFTPWYKKWVLYVNNYKKSTSEICHLHIIEP----------------------LKYNET------FELKPFQYS--LPDEFLQ--------YIPKSKWCLPDVSEEAALSRLKDFLG
NP_171935.1      KPFTSFNSYWKKCLDMSI---ESVM-LPPPWRLMP----------------------ITAAAEAIWA--CSIEELGL-----ENEAEKPS---N-ALLTRAW---SPGWSNADKLLNEFIE
YP_019820.1      TEYKVFTPFYNAFQKQVI--------HKPISKVQS----------------------IKGGNS--LPVSLSVSELHL---LPTIPWT--------SHMESIW---EPTEEGAYKTWKEFFS
NP_790955.1      NYFQVFTQFKKVCYTRLH---QAMPRTVHTPEAQQ-------------------PLSIKSDA------------------IPDQVKGFST---PSKTLRDLW---PAGEVEASRRIEAFAG
NP_001047200.1   CPFTMFAPFWDRCLCMP----DPAAPLLPPKRIAP------------------------GELP---ARRCPSDELVF-----EDESERGS---N-ALLARAW---SPGWQNADKALAAFLN
YP_793122.1      GYFQVFSQFRKVCHERLY---QALPGVRPRPQPQP-------------------PHALASDP------------------LPDAVPAFPR---PADSLRRLW---PAGEEVAQERLRGFAD
AC174468_14.1    QPFTTFDSFWERCLSMPY---DPQAPLLPPKRIIP------------------------GDVS---K--CPSDTLVF-----EDDLEKAS---N-ALLARAW---SPGWSNANKALTTFIN
NP_464116.1      EYYKIFTPYYKKWREEIK--------ETPFKVTLK-------------------ADNIRKENL-----------------FPKYEEQFAE---M-TCDLPIL---DSGEKTANTRLANFIK
NP_567341.1      RPFSMFAAFWERCLSMPY---DPESPLLPPKKIIS------------------------GDVS---K--CVADPLVF-----EDDSEKGS---N-ALLARAW---SPGWSNGDKALTTFIN
YP_029213.1      TEYKVFTPFYNAFQKQVI--------HKPISKVQS----------------------IKGGNS--LPVSLSVSELHL---LPTIPWT--------SHMESIW---EPTEEGAYKTWKEFFS
NP_820171.1      GYYKIFTPFWKHINASLQ----PRKTLAAPKKFSQ-------------------HRAIKSDN---------LETWNL---LPKHPDW--------SQGFDQW---KPGEKNAQIVLKKFIK
NP_718938.1      DMYRVFTPFSRKWREVAA-----KQAILPLGVPAP-------------------LGSILSEPKPLFFKNIGSTYLAASKDIASSKKGFNQTAPNPLISSEQW---AAGEGQAKRLLNQFIQ
NP_253349.1      GYFQVFSQFRKVCHERLY---QALPGVRPRPQPQP-------------------PHALASDP------------------LPDAVPAFPR---PADSLRRLW---PAGEEVAQERLRDFAD
NP_849588.1      KPFTSFNSYWKKCLDMSI---ESVM-LPPPWRLMP----------------------ITAAAEAIWA--CSIEELGL-----ENEAEKPS---N-ALLTRAW---SPGWSNADKLLNEFIE
AC122161_5.2     HAFTTFDPFWSRCLQMQM---KTYS-TIPPCQLIP----------------------AQGKVE---K--HSIEQLGL-----EDELETSS---N-ALLSRAW---SPGWSKSDKALTEFVE
NP_232458.1      EFFKVFTPFKRAWLTLFQ---PPVIGKNRPVALWN--------------------------------VPSALAELVW-----HPEQAFDY---P-RIDSTPW---AADFETVRAQLRDFCR
YP_013222.1      EYYKIFTPYYKKWREEIK--------ETPFKVTLK-------------------HENIRKESL-----------------FLKYEEQFAE---M-TSDLPIL---DIGERAANTRLANFIK
YP_167152.1      GFYKVYTPFWRAVSQREVSAPDPAPARVRPPEAWP--------------------------------GSEALSDWGLG--APMRRGA--------AVVATHQ---APGEAAALERLDRFIA
YP_234055.1      NYFQVFTQFKKVCYTRLH---EAMPHPVRTPQAQA-------------------AQSIKSDA------------------VPKQVKGFAT---PSKALRDLW---PAGEAEARRRMDAFAA
AC122171_26.2    HAFTTFDPFWSRCLQMQM---KTYS-TIPPCQLIP----------------------AQGKVE---K--HSIEQLGL-----EDELETSS---N-ALLSRAW---SPGWSKSDKALTEFVE
Q0GKU4_BRACM     RPFSMFAAFWERCLSMPY---DPESPLLPPKKIISGLLLHHTLSEIVSYPLLITLFVLAGDVS---K--CVADTLIF-----EDESEKGS---N-ALLARAW---SPGWSNADKALTTFIN
Q1JU52_BRANA     RPFSMFAAFWERCLSMPY---DPESPLLPPKKIIS------------------------GDVS---K--CVADTLIF-----EDESEKGS---N-ALLARAW---SPGWSNADKALTTFIN
Q309E8_NICSY     RPFTTFSAFWEKCLSMPY---DPEAPLLPPKRIIP------------------------GDAS---R--CSSDSLVF-----EDESEKGS---N-ALLARAW---SPGWSNADKALTTFVN
A7NUY5_VITVI     HSFTTFSAFWDRCLSMPY---DPEAPLLPPKRINP------------------------GDVS---R--CPSDTIAF-----EDESEKGS---N-ALLARAW---TPGWSNADKALTIFIN
Q9XHD8_SOLLC     RPFTTFSAFWEKCLSMPY---DPEAPLLPPKRIIS------------------------GDAS---R--CPSDNLVF-----EDESEKGS---N-ALLARAW---SPGWSNADKALTTFVN
Q6YBV9_PEA       QPFTTFDSFWERCLSMPY---DPQAPLLPPKRIIP------------------------GDVS---R--CPSDTLVF-----EDELEKSS---N-ALLARAW---SPGWSNANKALTTFIN
Q6EAN1_PEA       QPFTTFDSFWERCLSMPY---DPQAPLLPPKRIIP------------------------GDVS---R--CPSDTLVF-----EDELEKSS---N-ALLARAW---SPGWSNANKALTTFIN
Q93VS0_SOLLC     RPFSTFSDFWEKCLTMPY---DPEAPLLPPKRIIS------------------------GDAS---R--CPSDNLVF-----ESELEKGS---N-ALLARAW---SPGWSNADKALTTFIN
```

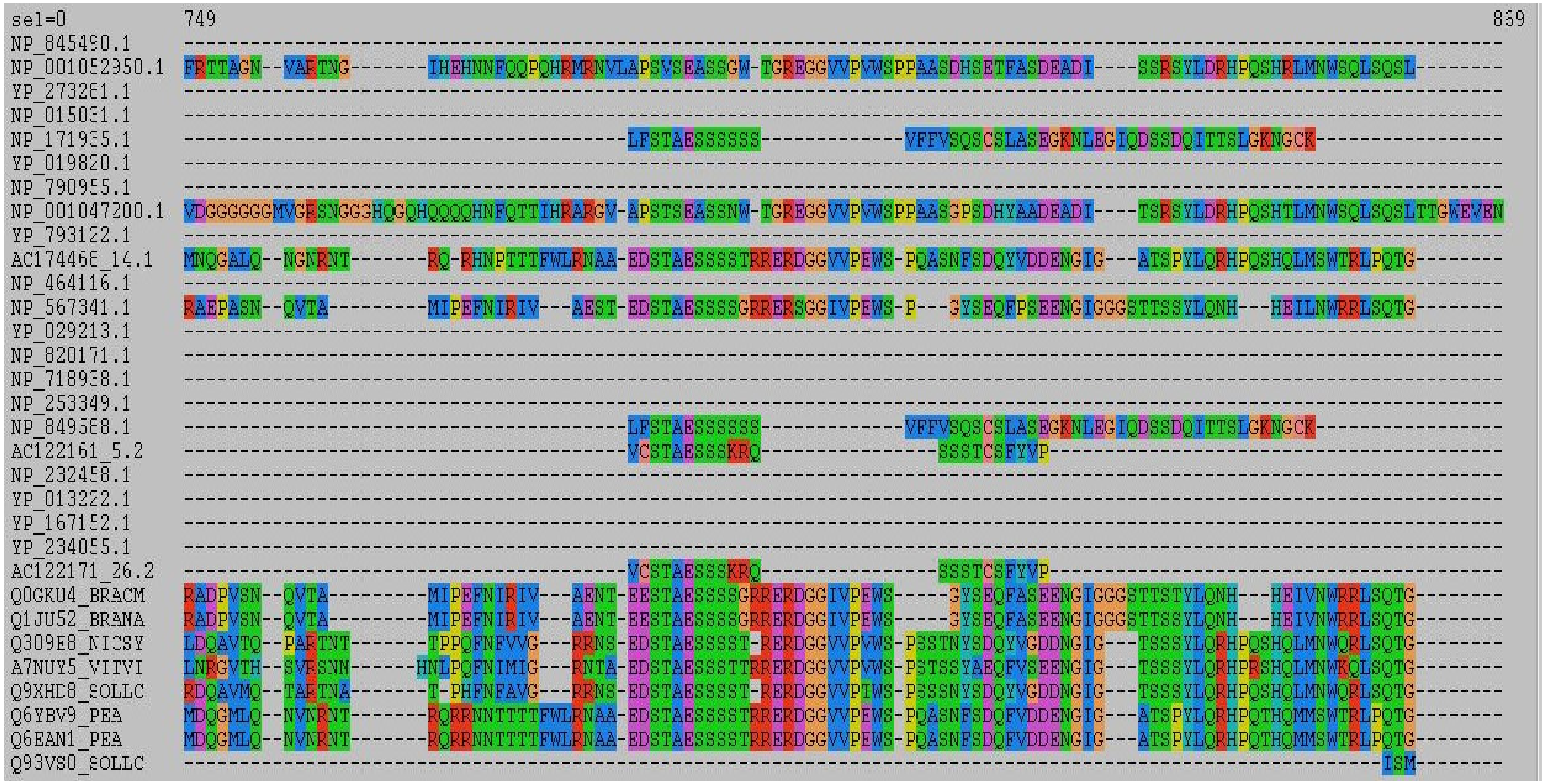

sel=0
361
481
NP_845490.1
NP_001052950.1
YP_273281.1
NP_015031.1
NP_171935.1
YP_019820.1
NP_790955.1
NP_001047200.1
YP_793122.1
ACT74468_14.1
NP_464116.1
NP_567341.1
YP_029213.1
NP_820171.1
NP_718938.1
NP_253349.1
NP_849588.1
ACI22161_5.2
NP_232458.1
YP_013222.1
YP_167152.1
YP_234055.1
ACI22171_26.2
Q0GKU4_BRACM
Q1JU52_BRANA
Q309E8_NICSY
A7NUY5_VITVI
Q9XHD8_SOLLC
Q6YBV9_PEA
Q6EAN1_PEA
Q93VS0_SOLLC


sel=0
NP_845490.1
NP_001052950.1
YP_273281.1
NP_015031.1
NP_171935.1
YP_019820.1
NP_790955.1
NP_001047200.1
YP_793122.1
AC174468_14.1
NP_464116.1
NP_567341.1
YP_029213.1
NP_820171.1
NP_718938.1
NP_253349.1
NP_849588.1
AC122161_5.2
NP_232458.1
YP_013222.1
YP_167152.1
YP_234055.1
AC122171_26.2
Q0GKU4_BRACM
Q1JU52_BRANA
Q309E8_NICSY
A7NUY5_VITVI
Q9XHD8_SOLLC
Q6YBV9_PEA
Q6EAN1_PEA
Q93VS0_SOLLC
481
601


sel=0
601
721
NP_845490.1
NP_001052950.1
YP_273281.1
NP_015031.1
NP_171935.1
YP_019820.1
NP_790955.1
NP_001047200.1
YP_793122.1
AC174468_14.1
NP_464116.1
NP_567341.1
YP_029213.1
NP_820171.1
NP_718938.1
NP_253349.1
NP_849588.1
AC122161_5.2
NP_232458.1
YP_013222.1
YP_167152.1
YP_234055.1
AC122171_26.2
Q0GKU4_BRACM
Q1JU52_BRANA
Q309E8_NICSY
A7NUY5_VITVI
Q9XHD8_SOLLC
Q6YBV9_PEA
Q6EAN1_PEA
Q93VS0_SOLLC



```
sel=0           721                                                                                                                   841
NP_845490.1     -------------------------------------------------------------------------------------------------------------------------
NP_001052950.1  DFMNSV-DSRAEVPTRVNFEPRTEREENFRTTAGN--VARTNG-------IHEHNNFQQPQHRMRNVLAPSVSEASSGW-TGREGGVVPVWSPPAASDHSETFASDEADI----SSRSYLD
YP_273281.1     -------------------------------------------------------------------------------------------------------------------------
NP_015031.1     -------------------------------------------------------------------------------------------------------------------------
NP_171935.1     KKSRGF-DER----------E-----------------------------------------------LFSTAESSSSSS-------------VFFVSQSCSLASEGKNLEGIQDSSDQIT
YP_019820.1     -------------------------------------------------------------------------------------------------------------------------
NP_790955.1     -------------------------------------------------------------------------------------------------------------------------
NP_001047200.1  DFDNSM-DSRPEVPSQVLFQPRMEREETVDGGGGGGMVGRSNGGGHQGQHQQQQHNFQTTIHRARGV-APSTSEASSNW-TGREGGVVPVWSPPAASGPSDHYAADEADI----TSRSYLD
YP_793122.1     -------------------------------------------------------------------------------------------------------------------------
AC174468_14.1   SVRNSAGDSRAEVPTN-A-----NAREAMNQGALQ--NGNRNT-------RQ-RHNPTTTFWLRNAA-EDSTAESSSSTRRERDGGVVPEWS-PQASNFSDQYVDDENGIG---ATSPYLQ
NP_464116.1     -------------------------------------------------------------------------------------------------------------------------
NP_567341.1     NLRNSVGDSRAEVPRNMVNTNQAQQR--RAEPASN--QVTA---------MIPEFNIRIV---AEST-EDSTAESSSSGRRERSGGIVPEWS-P---GYSEQFPSEENGIGGGSTTSSYLQ
YP_029213.1     -------------------------------------------------------------------------------------------------------------------------
NP_820171.1     -------------------------------------------------------------------------------------------------------------------------
NP_718938.1     -------------------------------------------------------------------------------------------------------------------------
NP_253349.1     -------------------------------------------------------------------------------------------------------------------------
NP_849588.1     KKSRGF-DER----------E-----------------------------------------------LFSTAESSSSSS-------------VFFVSQSCSLASEGKNLEGIQDSSDQIT
AC122161_5.2    DSLSNL-SKRTDTGVSCVDQE-----------------------------------------------VCSTAESSSKRQ----------------SSSTCSFYVP---------------
NP_232458.1     -------------------------------------------------------------------------------------------------------------------------
YP_013222.1     -------------------------------------------------------------------------------------------------------------------------
YP_167152.1     -------------------------------------------------------------------------------------------------------------------------
YP_234055.1     -------------------------------------------------------------------------------------------------------------------------
AC122171_26.2   DSLSNL-SKRTDTGVSCVDQE-----------------------------------------------VCSTAESSSKRQ----------------SSSTCSFYVP---------------
Q0GKU4_BRACM    SLRNSGGDSRAEVPRNMVNTNQARQQEARADPVSN--QVTA---------MIPEFNIRIV---AENT-EESTAESSSSGRRERDGGIVPEWS-----GYSEQFASEENGIGGGSTTSTYLQ
Q1JU52_BRANA    SLRNSGGDSRAEVPRNMVNTNQARQQEARADPVSN--QVTA---------MIPEFNIRIV---AENT-EESTAESSSSGRRERDGGIVPEWS-----GYSEQFASEENGIGGGSTTSSYLQ
Q309E8_NICSY    NIRNSVVDTRAEVPND-VNVTEGPRRDTLDQAVTQ--PARTNT-------TPPQFNFVVG---RRNS-EDSTAESSSST-RERDGGVVPVWS-PSSTNYSDQYVGDDNGIG---TSSSYLQ
A7NUY5_VITVI    DIQNSAENSRAEVPTN-VNANQEPRRETLNRGVTH--SVRSNN------HNLPQFNIMIG---RNTA-EDSTAESSSTTRRERDGGVVPVWS-PSTSSYAEQFVSEENGIG---TSSSYLQ
Q9XHD8_SOLLC    DIRNSVVESRAEVPTD-INVAEVHRRDTRDQAVMQ--TARTNA-------T-PHFNFAVG---RRNS-EDSTAESSSST-RERDGGVVPTWS-PSSSNYSDQYVGDDNGIG---TSSSYLQ
Q6YBV9_PEA      SVRNSAGDSRAEVPTN-AN-TQQNGREPMDQGMLQ--NVNRNT-------RQRRNNTTTTFWLRNAA-EDSTAESSSSTRRERDGGVVPEWS-PQASNFSDQFVDDENGIG---ATSPYLQ
Q6EAN1_PEA      SVRNSAGDSRAEVPTN-AN-TQQNGREPMDQGMLQ--NVNRNT-------RQRRNNTTTTFWLRNAA-EDSTAESSSSTRRERDGGVVPEWS-PQASNFSDQFVDDENGIG---ATSPYLQ
Q93VS0_SOLLC    DIRNSVVDSRAEVPI----------------------------------------------------------------------------------------------------------
```



```
sel=0            749                                                                                                                869
NP_845490.1      -----------------------------------------------------------------------------------------------------------------------
NP_001052950.1   FRTTAGN--VARTNG-------IHEHNNFQQPQHRMRNVLAPSVSEASSGW-TGREGGVVPVWSPPAASDHSETFASDEADI----SSRSYLDRHPQSHRLMNWSQLSQSL--------
YP_273281.1      -----------------------------------------------------------------------------------------------------------------------
NP_015031.1      -----------------------------------------------------------------------------------------------------------------------
NP_171935.1      ----------------------------------------LFSTAESSSSSS-------------VFFVSQSCSLASEGKNLEGIQDSSDQITTSLGKNGCK-----------------
YP_019820.1      -----------------------------------------------------------------------------------------------------------------------
NP_790955.1      -----------------------------------------------------------------------------------------------------------------------
NP_001047200.1   VDGGGGGGMVGRSNGGGHQGQHQQQQHNFQTTIHRARGV-APSTSEASSNW-TGREGGVVPVWSPPAASGPSDHYAADEADI----TSRSYLDRHPQSHTLMNWSQLSQSLTTGWEVEN
YP_793122.1      -----------------------------------------------------------------------------------------------------------------------
AC174468_14.1    MNQGALQ--NGNRNT-------RQ-RHNPTTTFWLRNAA-EDSTAESSSSTRRERDGGVVPEWS-PQASNFSDQYVDDENGIG---ATSPYLQRHPQSHQLMSWTRLPQTG--------
NP_464116.1      -----------------------------------------------------------------------------------------------------------------------
NP_567341.1      RAEPASN--QVTA---------MIPEFNIRIV---AEST-EDSTAESSSSGRRERSGGIVPEWS-P---GYSEQFPSEENGIGGGSTTSSYLQNH---HEILNWRRLSQTG--------
YP_029213.1      -----------------------------------------------------------------------------------------------------------------------
NP_820171.1      -----------------------------------------------------------------------------------------------------------------------
NP_718938.1      -----------------------------------------------------------------------------------------------------------------------
NP_253349.1      -----------------------------------------------------------------------------------------------------------------------
NP_849588.1      ----------------------------------------LFSTAESSSSSS-------------VFFVSQSCSLASEGKNLEGIQDSSDQITTSLGKNGCK-----------------
AC122161_5.2     ----------------------------------------VCSTAESSSKRQ----------------SSSTCSFYVP-----------------------------------------
NP_232458.1      -----------------------------------------------------------------------------------------------------------------------
YP_013222.1      -----------------------------------------------------------------------------------------------------------------------
YP_167152.1      -----------------------------------------------------------------------------------------------------------------------
YP_234055.1      -----------------------------------------------------------------------------------------------------------------------
AC122171_26.2    ----------------------------------------VCSTAESSSKRQ----------------SSSTCSFYVP-----------------------------------------
Q0GKU4_BRACM     RADPVSN--QVTA---------MIPEFNIRIV---AENT-EESTAESSSSGRRERDGGIVPEWS-----GYSEQFASEENGIGGGSTTSTYLQNH---HEIVNWRRLSQTG--------
Q1JU52_BRANA     RADPVSN--QVTA---------MIPEFNIRIV---AENT-EESTAESSSSGRRERDGGIVPEWS-----GYSEQFASEENGIGGGSTTSSYLQNH---HEIVNWRRLSQTG--------
Q309E8_NICSY     LDQAVTQ--PARTNT-------TPPQFNFVVG---RRNS-EDSTAESSSST-RERDGGVVPVWS-PSSTNYSDQYVGDDNGIG---TSSSYLQRHPQSHQLMNWQRLSQTG--------
A7NUY5_VITVI     LNRGVTH--SVRSNN------HNLPQFNIMIG---RNTA-EDSTAESSSTTRRERDGGVVPVWS-PSTSSYAEQFVSEENGIG---TSSSYLQRHPRSHQLMNWKQLSQTG--------
Q9XHD8_SOLLC     RDQAVMQ--TARTNA-------T-PHFNFAVG---RRNS-EDSTAESSSST-RERDGGVVPTWS-PSSSNYSDQYVGDDNGIG---TSSSYLQRHPQSHQLMNWQRLSQTG--------
Q6YBV9_PEA       MDQGMLQ--NVNRNT-------RQRRNNTTTTFWLRNAA-EDSTAESSSSTRRERDGGVVPEWS-PQASNFSDQFVDDENGIG---ATSPYLQRHPQTHQMMSWTRLPQTG--------
Q6EAN1_PEA       MDQGMLQ--NVNRNT-------RQRRNNTTTTFWLRNAA-EDSTAESSSSTRRERDGGVVPEWS-PQASNFSDQFVDDENGIG---ATSPYLQRHPQTHQMMSWTRLPQTG--------
Q93VS0_SOLLC     ------------------------------------------------------------------------------------------------------------ISM--------
```

XI. References

Ahmad, M., Jarillo, J.A. and Cashmore, A.R. (1998) Chimeric proteins between cry1 and cry2 Arabidopsis blue light photoreceptors indicate overlapping functions and varying protein stability, *Plant Cell*, **10**, 197-207.

Altschul, S.F. and Koonin, E.V. (1998) Iterated profile searches with PSI-Blast--a tool for discovery in protein databases, *Trends Biochem Sci*, **23**, 444-447.

Altschul, S., Madden, T., Schäfer, A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. (1997) Gapped Blast and PSI-Blast: a new generation of protein database search programs., *Nucleic Acids Res*, **25**, 3389-3402.

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J., Richardson, J., Ringwald, M., Rubin, G. and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium., *Nat Genet*, **25**, 25-29.

Bader, G.D., Betel, D. and Hogue, C.W. (2003) BIND: the Biomolecular Interaction Network Database, *Nucleic Acids Res*, **31**, 248-250.

Baldwin, G.S. (1993) Comparison of transferrin sequences from different species, *Comp Biochem Physiol B*, **106**, 203-218.

Ballvora, A., Ercolano, M., Weiss, J., Meksem, K., Bormann, C., Oberhagemann, P., Salamini, F. and Gebhardt, C. (2002) The R1 gene for potato resistance to late blight (Phytophthora infestans) belongs to the leucine zipper/NBS/LRR class of plant resistance genes., *Plant J*, **30**, 361-371.

Ballvora, A., Jocker, A., Viehover, P., Ishihara, H., Paal, J., Meksem, K., Bruggmann, R., Schoof, H., Weisshaar, B. and Gebhardt, C. (2007) Comparative sequence analysis of Solanum and Arabidopsis in a hot spot for pathogen resistance on potato chromosome V reveals a patchwork of conserved and rapidly evolving genome segments, *BMC Genomics*, **8**, 112.

Barone, A., Ritter, E., Schachtschabel, U., Debener, T., Salamini, F. and Gebhardt, C. (1990) Localization by restriction fragment length polymorphism mapping in potato of a major dominant gene conferring resistance to the potato cyst nematode Globodera rostochiensis., *Mol Gen Genet*, **224**, 177-182.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S., Griffiths-Jones, S., Howe, K., Marshall, M. and Sonnhammer, E. (2002) The Pfam protein families database., *Nucleic Acids Res*, **30**, 276-280.

Bendahmane, A., Querci, M., Kanyuka, K. and Baulcombe, D. (2000) Agrobacterium transient expression system as a tool for the isolation of disease resistance genes: application to the Rx2 locus in potato., *Plant J*, **21**, 73-81.

BioMoby, C., Wilkinson, M.D., Senger, M., Kawas, E., Bruskiewich, R., Gouzy, J., Noirot, C., Bardou, P., Ng, A., Haase, D., Saiz Ede, A., Wang, D., Gibbons, F., Gordon, P.M., Sensen, C.W., Carrasco, J.M., Fernandez, J.M., Shen, L., Links, M., Ng, M., Opushneva, N., Neerincx, P.B., Leunissen, J.A., Ernst, R., Twigger, S., Usadel, B., Good, B., Wong, Y., Stein, L., Crosby, W., Karlsson, J., Royo, R., Parraga, I., Ramirez, S., Gelpi, J.L., Trelles, O., Pisano, D.G., Jimenez, N., Kerhornou, A., Rosset, R., Zamacola, L., Tarraga, J., Huerta-Cepas, J., Carazo, J.M., Dopazo, J., Guigo, R., Navarro, A., Orozco, M., Valencia, A., Claros, M.G., Perez, A.J., Aldana, J., Rojano, M.M., Fernandez-Santa Cruz, R., Navas, I., Schiltz, G., Farmer, A., Gessler, D., Schoof, H. and Groscurth, A. (2008) Interoperability with Moby 1.0--it's better than sharing your toothbrush!, *Brief Bioinform*, **9**, 220-231.

Blumenthal, T. (2004) Operons in eukaryotes, *Brief Funct Genomic Proteomic*, **3**, 199-211.

Bonierbale, M., Plaisted, R. and Tanksley, S. (1988) RFLP Maps Based on a Common Set of Clones Reveal Modes of Chromosomal Evolution in Potato and Tomato., *Genetics*, **120**, 1095-1103.

Boutros, M., Agaisse, H. and Perrimon, N. (2002) Sequential activation of signaling pathways during innate immune responses in Drosophila, *Dev Cell*, **3**, 711-722.

Boutros, P.C. and Okey, A.B. (2005) Unsupervised pattern recognition: an introduction to the whys and wherefores of clustering microarray data, *Brief Bioinform*, **6**, 331-343.

Brenner, S.E., Chothia, C., Hubbard, T.J. and Murzin, A.G. (1996) Understanding protein structure: using scop for fold interpretation, *Methods Enzymol*, **266**, 635-643.

Brikos, C. and O'Neill, L.A. (2008) Signalling of toll-like receptors, *Handb Exp Pharmacol*, 21-50.

Brooks, J. and Wessel, G. (2002) The major yolk protein in sea urchins is a transferrin-like, iron binding protein., *Dev Biol*, **245**, 1-12.

Burns, D.M., Horn, V., Paluh, J. and Yanofsky, C. (1990) Evolution of the tryptophan synthetase of fungi. Analysis of experimentally fused Escherichia coli tryptophan synthetase alpha and beta chains, *J Biol Chem*, **265**, 2060-2069.

Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R. and Apweiler, R. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology, *Nucleic Acids Res*, **32**, D262-266.

Capriotti, E., Fariselli, P., Rossi, I. and Casadio, R. (2008) A three-state prediction of single point mutations on protein stability changes, *BMC Bioinformatics*, **9 Suppl 2**, S6.

Chan, A.P., Rabinowicz, P.D., Quackenbush, J., Buell, C.R. and Town, C.D. (2007) Plant database resources at the institute for genomic research, *Methods Mol Biol*, **406**, 113-136.

Chatterjee, M., Sharma, P. and Khurana, J.P. (2006) Cryptochrome 1 from Brassica napus is up-regulated by blue light and controls hypocotyl/stem growth and anthocyanin accumulation, *Plant Physiol*, **141**, 61-74.

Côé R., Jones, P., Martens, L., Kerrien, S., Reisinger, F., Lin, Q., Leinonen, R., Apweiler, R. and Hermjakob, H. (2007) The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases., *BMC Bioinformatics*, **8**, 401.

Date, S.V. (2008) The Rosetta stone method, *Methods Mol Biol*, **453**, 169-180.

de Castro, E., Sigrist, C.J., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P.S., Gasteiger, E., Bairoch, A. and Hulo, N. (2006) ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins, *Nucleic Acids Res*, **34**, W362-365.

De Jong, W., Forsyth, A., Leister, D., Gebhardt, C. and Baulcombe, D.C. (1997) A Potato hypersensitive resistance gene against potato virus X maps to a resistance gene cluster on chromosome V., *Theor. Appl. Genet.*, **95**, 153-162.

Delcher, A., Phillippy, A., Carlton, J. and Salzberg, S. (2002) Fast algorithms for large-scale genome alignment and comparison., *Nucleic Acids Res*, **30**, 2478-2483.

Duarte, J.M., Cui, L., Wall, P.K., Zhang, Q., Zhang, X., Leebens-Mack, J., Ma, H., Altman, N. and dePamphilis, C.W. (2006) Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of Arabidopsis, *Mol Biol Evol*, **23**, 469-478.

Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998) *Biological sequence analysis*. Cambridge University Press 1998.

Eckart, J.D. and Sobral, B.W. (2003) A life scientist's gateway to distributed data management and computing: the PathPort/ToolBus framework, *OMICS*, **7**, 79-88.

Edgar, R. (2004 (1)) MUSCLE: multiple sequence alignment with high accuracy and high throughput., *Nucleic Acids Res*, **32**, 1792-1797.

Emmert-Streib, F. and Dehmer, M. (2008) *Analysis of Microarray Data: A Network-Based Approach.* Wiley-VCH.

Enault, F., Suhre, K. and Claverie, J.M. (2005) Phydbac "Gene Function Predictor": a gene annotation tool based on genomic context analysis, *BMC Bioinformatics*, **6**, 247.

Engelhardt, B., Jordan, M. and Brenner, S. (2006) A graphical model for predicting protein molecular function. *Proceedings of the 23rd international conference on Machine learning*. ACM, Pittsburgh, Pennsylvania.

Engelhardt, B., Jordan, M., Muratore, K. and Brenner, S. (2005) Protein molecular function prediction by Bayesian phylogenomics., *PLoS Comput Biol*, **1**, e45.

Farjon, A. (1991) *Pinaceae: Drawings and Descriptions of the Genera : Abies, Cedrus, Pseudolarix, Keteleeria, Nothotsuga, Tsuga, Cathaya, Pseudotsuga, Larix and Picea.* Koeltz Scientific Books.

Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach, *J Mol Evol*, **17**, 368-376.

Felsenstein, J. (1993) PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author, Department of Genetics, University of Washington, Seattle.

Fisher, M., Gokhman, I., Pick, U. and Zamir, A. (1997) A structurally novel transferrin-like protein accumulates in the plasma membrane of the unicellular green alga Dunaliella salina grown in high salinities., *J Biol Chem*, **272**, 1565-1570.

Fisher, M., Zamir, A. and Pick, U. (1998) Iron uptake by the halotolerant alga Dunaliella is mediated by a plasma membrane transferrin., *J Biol Chem*, **273**, 17553-17558.

Fisher, P., Hedeler, C., Wolstencroft, K., Hulme, H., Noyes, H., Kemp, S., Stevens, R. and Brass, A. (2007) A systematic strategy for large-scale analysis of genotype phenotype correlations: identification of candidate genes involved in African trypanosomiasis, *Nucleic Acids Res*, **35**, 5625-5633.

Fleury, M. and Reverbel, F. (2003) *The JBoss Extensible Server.* Springer Berlin / Heidelberg.

Friedberg, I. (2006) Automated protein function prediction--the genomic challenge, *Brief Bioinform*, **7**, 225-242.

Galperin, M. and Koonin, E. (1998) Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption., *In Silico Biol*, **1**, 55-67.

Gardner, P.P., Daub, J., Tate, J.G., Nawrocki, E.P., Kolbe, D.L., Lindgreen, S., Wilkinson, A.C., Finn, R.D., Griffiths-Jones, S., Eddy, S.R. and Bateman, A. (2009) Rfam: updates to the RNA families database, *Nucleic Acids Res*, **37**, D136-140.

Gascuel, O. (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data., *Mol Biol Evol*, **14**, 685-695.

Gebhardt, C. (2004) Potato Genetics: Molecular Maps and More in Biotechnology in Agriculture and Forestry. Springer-Verlag, Berlin, Heidelberg.

Gebhardt, C., Ritter, E., Debener, T., Schachtschabel, U., Walkemeier, B., Uhrig, U. and Salamini, F. (1989) RFLP analysis and linkage mapping in Solanum tuberosum, **78**.

Gebhardt, C. and Valkonen, J. (2001) Organization of genes controlling disease resistance in the potato genome., *Annu Rev Phytopathol*, **39**, 79-102.

George, R.A., Spriggs, R.V., Thornton, J.M., Al-Lazikani, B. and Swindells, M.B. (2004) SCOPEC: a database of protein catalytic domains, *Bioinformatics*, **20 Suppl 1**, i130-136.

Gilks, W., Audit, B., De Angelis, D., Tsoka, S. and Ouzounis, C. (2002) Modeling the percolation of annotation errors in a database of protein sequences., *Bioinformatics*, **18**, 1641-1649.

Gish, W. (1996-2004) WU-Blast, http://Blast.wustl.edu.

Glanville, J.G., Kirshner, D., Krishnamurthy, N. and Sjolander, K. (2007) Berkeley Phylogenomics Group web servers: resources for structural phylogenomic analysis, *Nucleic Acids Res*, **35**, W27-32.

Glaser, F., Morris, R.J., Najmanovich, R.J., Laskowski, R.A. and Thornton, J.M. (2006) A method for localizing ligand binding pockets in protein structures, *Proteins*, **62**, 479-488.

Goble, C. and De Roure, D.C. (2007) *myExperiment: social networking for workflow-using e-scientists*. ACM.

Godzik, A., Jambon, M. and Friedberg, I. (2007) Computational protein function prediction: are we making progress?, *Cell Mol Life Sci*, **64**, 2505-2511.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H. and Oliver, S.G. (1996) Life with 6000 genes, *Science*, **274**, 546, 563-547.

Goff, S., Ricke, D., Lan, T., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., Hadley, D., Hutchison, D., Martin, C., Katagiri, F., Lange, B., Moughamer, T., Xia, Y., Budworth, P., Zhong, J., Miguel, T., Paszkowski, U., Zhang, S., Colbert, M., Sun, W., Chen, L., Cooper, B., Park, S., Wood, T., Mao, L., Quail, P., Wing, R., Dean, R., Yu, Y., Zharkikh, A., Shen, R., Sahasrabudhe, S., Thomas, A., Cannings, R., Gutin, A., Pruss, D., Reid, J., Tavtigian, S., Mitchell, J., Eldredge, G., Scholl, T., Miller, R., Bhatnagar, S., Adey, N., Rubano, T., Tusneem, N., Robinson, R., Feldhaus, J., Macalma, T., Oliphant, A. and Briggs, S. (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science*, **296**, 92-100.

Golovin, A. and Henrick, K. (2008) MSDmotif: exploring protein sites and motifs, *BMC Bioinformatics*, **9**, 312.

Gordon, D., Abajian, C. and Green, P. (1998) Consed: a graphical tool for sequence finishing, *Genome Res*, **8**, 195-202.

Gremme, G., Brendel, V., Sparks, M.E. and Kurtz, S. (2005) Engineering a software tool for gene structure prediction in higher organisms, *Information and Software Technology*, **47**, 965-978.

Grigoriev, A. (2003) On the number of protein-protein interactions in the yeast proteome, *Nucleic Acids Res*, **31**, 4157-4161.

Guindon, S. and Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood., *Syst Biol*, **52**, 696-704.

Hall, N. (2007) Advanced sequencing technologies and their wider impact in microbiology, *J Exp Biol*, **210**, 1518-1525.

Hand, D.J. and Heard, N.A. (2005) Finding groups in gene expression data, *J Biomed Biotechnol*, **2005**, 215-225.

Harrington, E.D., Jensen, L.J. and Bork, P. (2008) Predicting biological networks from genomic data, *FEBS Lett*, **582**, 1251-1258.

Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D. and Apweiler, R. (2004) IntAct: an open source molecular interaction database, *Nucleic Acids Res*, **32**, D452-455.

Holdener, A.T., St. Laurent, S. and Read, J. (2008) *Ajax. The Definitive Guide.* O'Reilly Media, Inc.

Horng, T., Barton, G.M. and Medzhitov, R. (2001) TIRAP: an adapter molecule in the Toll signaling pathway, *Nat Immunol*, **2**, 835-841.

Howe, K., Bateman, A. and Durbin, R. (2002) QuickTree: building huge Neighbour-Joining trees of protein sequences., *Bioinformatics*, **18**, 1546-1547.

Hsing, M., Byler, K.G. and Cherkasov, A. (2008) The use of Gene Ontology terms for predicting highly-connected 'hub' nodes in protein-protein interaction networks, *BMC Syst Biol*, **2**, 80.

Hsu, D.S., Zhao, X., Zhao, S., Kazantsev, A., Wang, R.P., Todo, T., Wei, Y.F. and Sancar, A. (1996) Putative human blue-light photoreceptors hCRY1 and hCRY2 are flavoproteins, *Biochemistry*, **35**, 13871-13877.

Huebers, H., Huebers, E., Finch, C., Webb, B., Truman, J., Riddiford, L., Martin, A. and Massover, W. (1988) Iron binding proteins and their roles in the tobacco hornworm, Manduca sexta (L.). *J Comp Physiol [B]*, **158**, 291-300.

Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., De Castro, E., Langendijk-Genevaux, P.S., Pagni, M. and Sigrist, C.J. (2006) The PROSITE database, *Nucleic Acids Res*, **34**, D227-230.

Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R.D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Mulder, N., Natale, D., Orengo, C., Quinn, A.F., Selengut, J.D., Sigrist, C.J., Thimma, M., Thomas, P.D., Valentin, F., Wilson, D., Wu, C.H. and Yeats, C. (2008) InterPro: the integrative protein signature database, *Nucleic Acids Res*.

Hurles, M. (2004) Gene duplication: the genomic trade in spare parts, *PLoS Biol*, **2**, E206.

European Bioinformatics Institute (2008) InterPro2GO
http://www.geneontology.org/external2go/interpro2go.

Irizarry, R.A., Warren, D., Spencer, F., Kim, I.F., Biswal, S., Frank, B.C., Gabrielson, E., Garcia, J.G., Geoghegan, J., Germino, G., Griffin, C., Hilmer, S.C., Hoffman, E., Jedlicka, A.E., Kawasaki, E., Martinez-Murillo, F., Morsberger, L., Lee, H., Petersen, D., Quackenbush, J., Scott, A., Wilson, M., Yang, Y., Ye, S.Q. and Yu, W. (2005) Multiple-laboratory comparison of microarray platforms, *Nat Methods*, **2**, 345-350.

Ivanisenko, V.A., Pintus, S.S., Grigorovich, D.A. and Kolchanov, N.A. (2004) PDBSiteScan: a program for searching for active, binding and posttranslational modification sites in the 3D structures of proteins, *Nucleic Acids Res*, **32**, W549-554.

Iwata, S., Lee, J.W., Okada, K., Lee, J.K., Iwata, M., Rasmussen, B., Link, T.A., Ramaswamy, S. and Jap, B.K. (1998) Complete structure of the 11-subunit bovine mitochondrial cytochrome bc1 complex, *Science*, **281**, 64-71.

JA, E. (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis., *Genome Res*, **8**, 163-167.

Jarvinen, A.K., Hautaniemi, S., Edgren, H., Auvinen, P., Saarela, J., Kallioniemi, O.P. and Monni, O. (2004) Are data from different gene expression microarray platforms comparable?, *Genomics*, **83**, 1164-1168.

Jauch, R., Yeo, H.C., Kolatkar, P.R. and Clarke, N.D. (2007) Assessment of CASP7 structure predictions for template free targets, *Proteins*, **69 Suppl 8**, 57-67.

Jöcker, A., Hoffmann, F., Groscurth, A. and Schoof, H. (2008) Protein function prediction and annotation in an integrated environment powered by web services (AFAWE). *Bioinformatics*, **24**, 2393-2394.

Jöcker, A., Jöcker, A., Engelhardt B.E., Göbel, U. and Schoof, H. (2009) Using Additional Functional Attributes in a Bayesian Phylogenomics approach for improving Function Prediction, in preparation.

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S. and Madden, T.L. (2008) NCBI Blast: a better web interface, *Nucleic Acids Res*, **36**, W5-9.

Kanai, S., Kikuno, R., Toh, H., Ryo, H. and Todo, T. (1997) Molecular evolution of the photolyase-blue-light photoreceptor family, *J Mol Evol*, **45**, 535-548.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T. and Yamanishi, Y. (2008) KEGG for linking genomes to life and the environment, *Nucleic Acids Res*, **36**, D480-484.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M. (2004) The KEGG resource for deciphering the genome., *Nucleic Acids Res*, **32**, D277-280.

Karimpour-Fard, A., Leach, S.M., Gill, R.T. and Hunter, L.E. (2008) Predicting protein linkages in bacteria: which method is best depends on task, *BMC Bioinformatics*, **9**, 397.

Karplus, K., Barrett, C. and Hughey, R. (1998) Hidden Markov models for detecting remote protein homologies, *Bioinformatics*, **14**, 846-856.

Katoh, K., Kuma, K., Toh, H. and Miyata, T. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment., *Nucleic Acids Res*, **33**, 511-518.

Kerr, G., Ruskin, H.J., Crane, M. and Doolan, P. (2008) Techniques for clustering gene expression data, *Comput Biol Med*, **38**, 283-293.

Khatri, P. and Draghici, S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems, *Bioinformatics*, **21**, 3587-3595.

Kinoshita, K., Furui, J. and Nakamura, H. (2002) Identification of protein functions from a molecular surface database, eF-site, *J Struct Funct Genomics*, **2**, 9-22.

Kirchmair, J., Markt, P., Distinto, S., Schuster, D., Spitzer, G.M., Liedl, K.R., Langer, T. and Wolber, G. (2008) The Protein Data Bank (PDB), Its Related Services and Software Tools as Key Components for In Silico Guided Drug Discovery, *J Med Chem*.

Kolesov, G., Mewes, H.W. and Frishman, D. (2001) SNAPping up functionally related genes based on context information: a colinearity-free approach, *J Mol Biol*, **311**, 639-656.

Kourmpetis, Y.A., van der Burgt, A., Bink, M.C., Ter Braak, C.J. and van Ham, R.C. (2007) The use of multiple hierarchically independent gene ontology terms in gene function prediction and genome annotation, *In Silico Biol*, **7**, 575-582.

Krawetz, S.A. and Womble, D.D. (2003) *Introduction to Bioinformatics: A Theoretical and Practical Approach*. Humana Press.

Kreike, C.M., De Koning, J.R.A., Vinke, J.H., Van Ooijen, J.W. and Stiekema, W.J. (1994) Quantitatively-inherited resistance to Globodera pallida is dominated by one major locus in Solanum spegazzinii., *Theor. Appl. Genet.*, **88**, 764-769.

Krishnamurthy, N., Brown, D. and Sjolander, K. (2007) FlowerPower: clustering proteins into domain architecture classes for phylogenomic inference of protein function, *BMC Evol Biol*, **7 Suppl 1**, S12.

Kriventseva, E.V., Koch, I., Apweiler, R., Vingron, M., Bork, P., Gelfand, M.S. and Sunyaev, S. (2003) Increase of functional diversity by alternative splicing, *Trends Genet*, **19**, 124-128.

Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes, *J Mol Biol*, **305**, 567-580.

Kuang, H., Wei, F., Marano, M., Wirtz, U., Wang, X., Liu, J., Shum, W., Zaborsky, J., Tallon, L., Rensink, W., Lobst, S., Zhang, P., Tornqvist, C., Tek, A., Bamberg, J., Helgeson, J., Fry, W., You, F., Luo, M., Jiang, J., Robin Buell, C. and Baker, B. (2005) The R1 resistance gene cluster contains three groups of independently evolving, type I R1 homologues and shows substantial structural variation among haplotypes of Solanum demissum., *Plant J*, **44**, 37-51.

Kulikova, T., Aldebert, P., Althorpe, N., Baker, W., Bates, K., Browne, P., van den Broek, A., Cochrane, G., Duggan, K., Eberhardt, R., Faruque, N., Garcia-Pastor, M., Harte, N., Kanz, C., Leinonen, R., Lin, Q., Lombard, V., Lopez, R., Mancuso, R., McHale, M., Nardone, F., Silventoinen, V., Stoehr, P., Stoesser, G., Tuli, M.A., Tzouvara, K., Vaughan, R., Wu, D., Zhu, W. and Apweiler, R. (2004) The EMBL Nucleotide Sequence Database, *Nucleic Acids Res*, **32**, D27-30.

Kurama, T., Kurata, S. and Natori, S. (1995) Molecular characterization of an insect transferrin and its selective incorporation into eggs during oogenesis., *Eur J Biochem*, **228**, 229-235.

Kurisu, G., Zhang, H., Smith, J.L. and Cramer, W.A. (2003) Structure of the cytochrome b6f complex of oxygenic photosynthesis: tuning the cavity, *Science*, **302**, 1009-1014.

Labarga, A., Valentin, F., Anderson, M. and Lopez, R. (2007) Web services at the European bioinformatics institute, *Nucleic Acids Res*, **35**, W6-11.

Lambert, L., Perri, H. and Meehan, T. (2005) Evolution of duplications in the transferrin family of proteins., *Comp Biochem Physiol B Biochem Mol Biol*, **140**, 11-25.

Lee, D., Redfern, O. and Orengo, C. (2007) Predicting protein function from sequence and structure, *Nat Rev Mol Cell Biol*, **8**, 995-1005.

Lee, Y. and Quackenbush, J. (2003) Using the TIGR gene index databases for biological discovery, *Curr Protoc Bioinformatics*, **Chapter 1**, Unit 1 6.

Leonards-Schippers, C., Gieffers, W., Salamini, F. and Gebhardt, C. (1992) The R1 gene conferring race-specific resistance to Phytophthora infestans in potato is located on potato chromosome V., *Mol Gen Genet*, **233**, 278-283.

Li, H., Coghlan, A., Ruan, J., Coin, L.J., Heriche, J.K., Osmotherly, L., Li, R., Liu, T., Zhang, Z., Bolund, L., Wong, G.K., Zheng, W., Dehal, P., Wang, J. and Durbin, R. (2006) TreeFam: a curated database of phylogenetic trees of animal gene families, *Nucleic Acids Res*, **34**, D572-580.

Lukashin, A. and Borodovsky, M. (1998) GeneMark.hmm: new solutions for gene finding., *Nucleic Acids Res*, **26**, 1107-1115.

Malhotra, K., Kim, S.T., Batschauer, A., Dawut, L. and Sancar, A. (1995) Putative blue-light photoreceptors from Arabidopsis thaliana and Sinapis alba with a high degree of sequence homology to DNA photolyase contain the two photolyase cofactors but lack DNA repair activity, *Biochemistry*, **34**, 6892-6899.

Marchler-Bauer, A., Anderson, J.B., Derbyshire, M.K., DeWeese-Scott, C., Gonzales, N.R., Gwadz, M., Hao, L., He, S., Hurwitz, D.I., Jackson, J.D., Ke, Z., Krylov, D., Lanczycki, C.J., Liebert, C.A., Liu, C., Lu, F., Lu, S., Marchler, G.H., Mullokandov, M., Song, J.S., Thanki, N., Yamashita, R.A., Yin, J.J., Zhang, D. and Bryant, S.H. (2007) CDD: a conserved domain database for interactive domain family analysis, *Nucleic Acids Res*, **35**, D237-240.

Marchler-Bauer, A., Panchenko, A.R., Shoemaker, B.A., Thiessen, P.A., Geer, L.Y. and Bryant, S.H. (2002) CDD: a database of conserved domain alignments with links to domain three-dimensional structure, *Nucleic Acids Res*, **30**, 281-283.

Martin, M.U. and Wesche, H. (2002) Summary and comparison of the signaling mechanisms of the Toll/interleukin-1 receptor family, *Biochim Biophys Acta*, **1592**, 265-280.

Meksem, K., Zobrist, K., Ruben, E., Hyten, D., Quanzhou, T., Zhang, H.-B. and Lightfoot, D.A. (2000) Two large-insert soybean genomic libraries constructed in a binary vector: applications in chromosome walking and genome wide physical mapping., *Theor. Appl. Genet.*, **101**, 747-755.

Misra, S. and Harris, N. (2006) Using Apollo to browse and edit genome annotations., *Curr Protoc Bioinformatics*, **Chapter 9**, Unit 9.5.

Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C. and Morris, Q. (2008) GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function, *Genome Biol*, **9 Suppl 1**, S4.

Mueller, L.A., Solow, T.H., Taylor, N., Skwarecki, B., Buels, R., Binns, J., Lin, C., Wright, M.H., Ahrens, R., Wang, Y., Herbst, E.V., Keyder, E.R., Menda, N., Zamir, D. and Tanksley, S.D. (2005) The SOL Genomics Network: a comparative resource for Solanaceae biology and beyond, *Plant Physiol*, **138**, 1310-1317.

Mulder, N. and Apweiler, R. (2007) InterPro and InterProScan: tools for protein sequence classification and comparison, *Methods Mol Biol*, **396**, 59-70.

Neerincx, P.B. and Leunissen, J.A. (2005) Evolution of web services in bioinformatics, *Brief Bioinform*, **6**, 178-188.

Ng, P.C. and Henikoff, S. (2006) Predicting the effects of amino acid substitutions on protein function, *Annu Rev Genomics Hum Genet*, **7**, 61-80.

Ninu, L., Ahmad, M., Miarelli, C., Cashmore, A.R. and Giuliano, G. (1999) Cryptochrome 1 controls tomato development in response to blue light, *Plant J*, **18**, 551-556.

Nuin, P.A., Wang, Z. and Tillier, E.R. (2006) The accuracy of several multiple sequence alignment programs for proteins, *BMC Bioinformatics*, **7**, 471.

O'Brien, K.P., Remm, M. and Sonnhammer, E.L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs, *Nucleic Acids Res*, **33**, D476-480.

Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Glover, K., Pocock, M.R., Wipat, A. and Li, P. (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows, *Bioinformatics*, **20**, 3045-3054.

Ong, S.T., Ho, J.Z., Ho, B. and Ding, J.L. (2006) Iron-withholding strategy in innate immunity, *Immunobiology*, **211**, 295-314.

Gene Ontology (2008) Gene Ontology association files,
http://www.geneontology.org/GO.current.annotations.shtml.

Pal, D. and Eisenberg, D. (2005) Inference of protein function from protein structure, *Structure*, **13**, 121-130.

Palmer Jd, S.D.E.C.M.W. (2004) The plant tree of life: An overview and some points of view, *American Journal of Botany*, **91**, 1437-1445.

Park, I., Schaeffer, E., Sidoli, A., Baralle, F.E., Cohen, G.N. and Zakin, M.M. (1985) Organization of the human transferrin gene: direct evidence that it originated by gene duplication, *Proc Natl Acad Sci U S A*, **82**, 3149-3153.

Pazos, F. and Sternberg, M.J. (2004) Automated prediction of protein function and detection of functional sites from structure, *Proc Natl Acad Sci U S A*, **101**, 14754-14759.

Perrodou, E., Chica, C., Poch, O., Gibson, T.J. and Thompson, J.D. (2008) A new protein linear motif benchmark for multiple sequence alignment software, *BMC Bioinformatics*, **9**, 213.

Perrotta, G., Ninu, L., Flamma, F., Weller, J.L., Kendrick, R.E., Nebuloso, E. and Giuliano, G. (2000) Tomato contains homologues of Arabidopsis cryptochromes 1 and 2, *Plant Mol Biol*, **42**, 765-773.

Platten, J.D., Foo, E., Foucher, F., Hecht, V., Reid, J.B. and Weller, J.L. (2005) The cryptochrome gene family in pea includes two differentially expressed CRY2 genes, *Plant Mol Biol*, **59**, 683-696.

Porter, C.T., Bartlett, G.J. and Thornton, J.M. (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data, *Nucleic Acids Res*, **32**, D129-133.

Poyatos, J.F. and Hurst, L.D. (2007) The determinants of gene order conservation in yeasts, *Genome Biol*, **8**, R233.

Presgraves, D.C. (2005) Evolutionary genomics: new genes for new jobs, *Curr Biol*, **15**, R52-53.

Price, M.N., Dehal, P.S. and Arkin, A.P. (2008) FastBlast: homology relationships for millions of proteins, *PLoS ONE*, **3**, e3589.

Pruitt, K., Tatusova, T., Klimke, W. and Maglott, D. (2008) NCBI Reference Sequences: current status, policy and new initiatives., *Nucleic Acids Res*.

Repsys, V., Margelevicius, M. and Venclovas, C. (2008) Re-searcher: a system for recurrent detection of homologous protein sequences, *BMC Bioinformatics*, **9**, 296.

Rice, P., Longden, I. and Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite, *Trends Genet*, **16**, 276-277.

Rost, B. (1997) Protein structures sustain evolutionary drift, *Fold Des*, **2**, S19-24.

Rost, B. (2002) Enzyme function less conserved than anticipated, *J Mol Biol*, **318**, 595-608.

Ruan, J., Li, H., Chen, Z., Coghlan, A., Coin, L.J., Guo, Y., Heriche, J.K., Hu, Y., Kristiansen, K., Li, R., Liu, T., Moses, A., Qin, J., Vang, S., Vilella, A.J., Ureta-Vidal, A., Bolund, L., Wang, J. and Durbin, R. (2008) TreeFam: 2008 Update, *Nucleic Acids Res*, **36**, D735-740.

Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., Tetko, I., Guldener, U., Mannhaupt, G., Munsterkotter, M. and Mewes, H.W. (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes, *Nucleic Acids Res*, **32**, 5539-5545.

Salamov, A.A. and Solovyev, V.V. (2000) Ab initio gene finding in Drosophila genomic DNA, *Genome Res*, **10**, 516-522.

Salgado, H., Moreno-Hagelsieb, G., Smith, T.F. and Collado-Vides, J. (2000) Operons in Escherichia coli: genomic analyses and predictions, *Proc Natl Acad Sci U S A*, **97**, 6652-6657.

Sancar, A. (2003) Structure and function of DNA photolyase and cryptochrome blue-light photoreceptors, *Chem Rev*, **103**, 2203-2237.

Sang, Y., Li, Q.H., Rubio, V., Zhang, Y.C., Mao, J., Deng, X.W. and Yang, H.Q. (2005) N-terminal domain-mediated homodimerization is required for photoreceptor activity of Arabidopsis CRYPTOCHROME 1, *Plant Cell*, **17**, 1569-1584.

Sato, Y., Nakaya, A., Shiraishi, K., Kawashima, S., Goto, S. and Kanehisa, M. (2001) SSDB: sequence similarity database in KEGG., *Genome Informatics*, **12**, 230-231.

Sayers, E.W., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., Dicuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L.Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D.J., Madden, T.L., Maglott, D.R., Miller, V., Mizrachi, I., Ostell, J., Pruitt, K.D., Schuler, G.D., Sequeira, E., Sherry, S.T., Shumway, M., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusova, T.A., Wagner, L., Yaschenko, E. and Ye, J. (2008) Database resources of the National Center for Biotechnology Information, *Nucleic Acids Res*.

Schiex, T., Moisan, A. and Rouzé P. (2001) EuGene: An Eucaryotic Gene Finder that combines several sources of evidence., 111-125.

Schulze-Lefert, P. (2004) Plant immunity: the origami of receptor activation, *Curr Biol*, **14**, R22-24.

Shirasu, K. and Schulze-Lefert, P. (2000) Regulators of cell death in disease resistance, *Plant Mol Biol*, **44**, 371-385.

Spannagl, M., Noubibou, O., Haase, D., Yang, L., Gundlach, H., Hindemitt, T., Klee, K., Haberer, G., Schoof, H. and Mayer, K. (2007) MIPSPlantsDB--plant database resource for integrative and comparative plant genome research., *Nucleic Acids Res*, **35**, D834-840.

Standley, D.M., Kinjo, A.R., Kinoshita, K. and Nakamura, H. (2008) Protein structure databases with new web services for structural biology and biomedical research, *Brief Bioinform*, **9**, 276-285.

Teichmann, S.A. and Babu, M.M. (2002) Conservation of gene co-regulation in prokaryotes and eukaryotes, *Trends Biotechnol*, **20**, 407-410; discussion 410.

Teichmann, S.A. and Veitia, R.A. (2004) Genes encoding subunits of stable complexes are clustered on the yeast chromosomes: an interpretation from a dosage balance perspective, *Genetics*, **167**, 2121-2125.

Thibaud-Nissen, F., Campbell, M., Hamilton, J.P., Zhu, W. and Buell, C.R. (2007) EuCAP, a Eukaryotic Community Annotation Package, and its application to the rice genome, *BMC Genomics*, **8**, 388.

Thimm, O., Blasing, O., Gibon, Y., Nagel, A., Meyer, S., Kruger, P., Selbig, J., Muller, L.A., Rhee, S.Y. and Stitt, M. (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes, *Plant J*, **37**, 914-939.

Thomas, P.D., Mi, H. and Lewis, S. (2007) Ontology annotation: mapping genomic regions to biological function, *Curr Opin Chem Biol*, **11**, 4-11.

Thompson, G.J., Crozier, Y.C. and Crozier, R.H. (2003) Isolation and characterization of a termite transferrin gene up-regulated on infection, *Insect Mol Biol*, **12**, 1-7.

Troyanskaya, O.G., Dolinski, K., Owen, A.B., Altman, R.B. and Botstein, D. (2003) A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae), *Proc Natl Acad Sci U S A*, **100**, 8348-8353.

UniProt, C. (2007) The Universal Protein Resource (UniProt), *Nucleic Acids Res*, **35**, D193-197.

Valles, S.M. and Pereira, R.M. (2005) Solenopsis invicta transferrin: cDNA cloning, gene architecture, and up-regulation in response to *Beauveria bassiana* infection, *Gene*, **358**, 60-66.

von Mering, C., Jensen, L.J., Kuhn, M., Chaffron, S., Doerks, T., Kruger, B., Snel, B. and Bork, P. (2007) STRING 7--recent developments in the integration and prediction of protein interactions, *Nucleic Acids Res*, **35**, D358-362.

Wapinski, I., Pfeffer, A., Friedman, N. and Regev, A. (2007) Automatic genome-wide reconstruction of phylogenetic gene trees, *Bioinformatics*, **23**, i549-558.

Webb, E.C. and NC-ICBMB. (1992) *Enzyme nomenclature 1992 : recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes [WorldCat.org]*. Academic Press Inc..

Weems, D., Miller, N., Garcia-Hernandez, M., Huala, E. and Rhee, S.Y. (2004) Design, Implementation and Maintenance of a Model Organism Database for Arabidopsis thaliana, *Comp Funct Genomics*, **5**, 362-369.

Wei, Y., Ko, J., Murga, L.F. and Ondrechen, M.J. (2007) Selective prediction of interaction sites in protein structures with THEMATICS, *BMC Bioinformatics*, **8**, 119.

Xiao, G. and Pan, W. (2005) Gene function prediction by a combined analysis of gene expression data and protein-protein interaction data, *J Bioinform Comput Biol*, **3**, 1371-1389.

Yendrek, C.R. and Metzger, J.D. (2005) Investigating the physiological effects of altered cryptochrome levels in the long day plant *Nicotiana sylvestris*.

Zhao, X.M., Wang, Y., Chen, L. and Aihara, K. (2008) Gene function prediction using labeled and unlabeled data, *BMC Bioinformatics*, **9**, 57.

Zmasek, C. and Eddy, S. (2001) A simple algorithm to infer gene duplication and speciation events on a gene tree., *Bioinformatics*, **17**, 821-828.

Zmasek, C. and Eddy, S. (2001) ATV: display and manipulation of annotated phylogenetic trees., *Bioinformatics*, **17**, 383-384.

List of Publications

Jöcker, A., Hoffmann, F., Groscurth, A. and Schoof, H. (2008) Protein function prediction and annotation in an integrated environment powered by web services (AFAWE). *Bioinformatics*, **24**, 2393-2394.

Ballvora, A., Jöcker, A., Viehover, P., Ishihara, H., Paal, J., Meksem, K., Bruggmann, R., Schoof, H., Weisshaar, B. and Gebhardt, C. (2007) Comparative sequence analysis of Solanum and Arabidopsis in a hot spot for pathogen resistance on potato chromosome V reveals a patchwork of conserved and rapidly evolving genome segments, *BMC Genomics*, **8**, 112.

Jöcker, A., Jöcker, A., Engelhardt, B.E., Göbel, U., Schoof, H. (2009) Using Additional Functional Attributes in a Bayesian Phylogenomics approach for improving Function Prediction, in preparation.

Declaration

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie – abgesehen von unten angegebenen Teilpublikationen – noch nicht veröffentlicht worden ist sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Dr. Thomas Wiehe betreut worden.

Köln, den 11 Februar 2009,

Curriculum vitae

Persöhnliche Daten

Name	Anika Jöcker
Geburtsdatum	23.10.1980
Geburtsort	Haan
Familienstand	verheiratet
Staatsangehörigkeit	Deutsch

Hochschulbildung

1991 – 2000
Wilhelm-Dörpfeld-Gymnasium in Wuppertal

2000 – 2006
Studium der Naturwissenschaftlichen Informatik an der Universität Bielefeld

2005
Diplomarbeit bei der BASF AG

Thema: Annotation und metabolische Rekonstruktion von *Magnaporthe grisea* als Modellorganismus für filamentöse Ascomycota.

Arbeitsverhältnisse

2002 – 2003
Studentische Hilfskraft in der Neuroinformatik Arbeitsgruppe von Prof. Dr. Ritter an der Universität Bielefeld.

Januar – Juni 2004
Studentische Hilfekraft in der Arbeitsgruppe Bioinformatik von Prof. Dr. Pühler an der Universität Bielefeld.

Juli - Oktober 2004
Praktikum bei der Firma Miele in Gütersloh.

Entwicklung einer datenbankgestützten Webanwendung unter Verwendung der Java-Servlets- und Java–Server-Pages Technologie.

März 2005
Praktikum bei der BASF AG in Ludwigshafen

Dissertation

Dezember 2005 – Februar 2009
Max-Planck Institut für Züchtungsforschung, unabhängige Forschungsgruppe "Plant Computational Biology" von Dr. Heiko Schoof. Betreut wurde die Arbeit von Prof. Dr. Thomas Wiehe vom Institut für Genetik an der Universität zu Köln