Automated structural annotation of the malaria proteome and identification of candidate proteins for modelling and crystallization studies

by

YOLANDI JOUBERT

Submitted in partial fulfillment of the requirements for the degree

Magister Scientiae

In the Faculty of Natural and Agricultural Sciences

Department of Biochemistry

University of Pretoria

Pretoria

June 2007

Declaration

I declare that the dissertation that I hereby submit for the degree in Bioinformatics at the University of Pretoria is my own work and that it has not previously been submitted by me for degree purposes at any other university.

Yolandi Joubert

Summary

Malaria is the cause of over one million deaths per year, primarily in African children. The parasite responsible for the most virulent form of malaria, is *Plasmodium falciparum*. Protein structure plays a pivotal role in elucidating mechanisms of parasite functioning and resistance to anti-malarial drugs. Protein structure furthermore aids the determination of protein function, which can together with the structure be used to identify novel drug targets in the parasite. However, various structural features in *P. falciparum* proteins complicate the experimental determination of protein three dimensional structures. Furthermore, the presence of parasite-specific inserts results in reduced similarity of these proteins to orthologous proteins with experimentally determined structures. The lack of solved structures in the malaria parasite, together with limited similarities to proteins in the Protein Data Bank, necessitate genome-scale structural annotation of *P. falciparum* proteins. Additionally, the annotation of a range of structural features facilitates the identification of suitable targets for structural studies.

An integrated structural annotation system was constructed and applied to all the predicted proteins in *P. falciparum*, *Plasmodium vivax* and *Plasmodium yoelii*. Similarity searches against the PDB, Pfam, Superfamily, PROSITE and PRINTS were included. In addition, the following predictions were made for the *P. falciparum* proteins: secondary structure, transmembrane helices, protein disorder, low complexity, coiled-coils and small molecule interactions. *P. falciparum* protein-protein interactions and proteins exported to the RBC were annotated from literature. Finally, a selection of proteins were threaded through a library of SCOP folds. All the results are stored in a relational PostgreSQL database and can be viewed through a web interface (`http://deepthought.bi.up.ac.za:8080/Annotation`). In order to select groups of proteins which fulfill certain criteria with regard to structural and functional features, a query tool was constructed. Using this tool, criteria regarding the presence or absence of all the predicted features can be specified.

Analysis of the results obtained revealed that *P. falciparum* protein-interacting proteins contain a higher percentage of predicted disordered residues than non-interacting proteins. Proteins interacting with 10 or more proteins have a disordered content concentrated in the range of 60-100%, while the disorder distribution for proteins having only one interacting partner, was more evenly spread. Comparisons of structural and sequence features between the three species, revealed that *P. falciparum* proteins tend to be longer and vary more in length than the other two species. *P. falciparum* proteins also contained more predicted low complexity and disorder content than proteins from *P. yoelii* and *P. vivax*.

P. falciparum protein targets for experimental structure determination, comparative modeling and *in silico* docking studies were putatively identified based on structural features. For experimental structure determination, 178 targets were identified. These targets contain limited contents of predicted transmembrane helix, disorder, coiled-coils, low complexity and signal peptide, as these features may complicate steps in the experimental structure determination procedure. In addition, the targets display low similarity to proteins in the PDB. Comparisons of the targets to proteins with crystal structures, revealed that the structures and predicted targets had similar sequence properties and predicted structural features. A group of 373 proteins which displayed high levels of similarity to proteins in the PDB, were identified as targets for comparitive modeling studies. Finally, 197 targets for *in silico* docking were identified based on predicted small molecule interactions and the availability of a 3D structure.

Acknowledgments

I would like to express my gratitude to the following people, organizations and institutes for assisting me in the completion of this project:

- To Prof. Fourie Joubert, for his professional, creative and insightful guidance of this project and for his generous support.
- To my parents and my sister for providing unconditional support and love during the course of my MSc.
- To John for all his love, encouragement and support.
- To all my past and present colleagues in the Bioinformatics and Computational Biology Unit of the University of Pretoria, especially to Ayton, Charles, Gordon and Hamilton.
- To all additional research colleagues and friends met in the duration of this project, for their encouragement and support.
- To the Department of Biochemistry and the Bioinformatics and Computational Biology Unit of the University of Pretoria for providing facilities and a sound academic environment.
- To the National Research Foundation of South Africa (NRF) and the University of Pretoria for awarding me bursaries which enabled me to undertake an MSc degree.
- To the International Society for Computational Biology (ISCB), Whitehead Scientific and the NRF for supplying funding for the attendance of the 13th and 14th annual meetings on Intelligent Systems for Molecular Biology (ISMB).

Contents

List of Figures

List of Tables

List of Abbreviations

3D Three dimensional

ANN Artificial neural network

ARW Average residue weight

EC Extinction coefficient

ER Endoplasmic reticulem

HMM Hidden Markov Model

IEIB Improbability of expression in inclusion bodies

IP Iso-electric point

MEC Molar extinction coefficient

MW Molecular weight

NMR Nuclear magnetic resonance

PC Priority class

PDE Phosphodiesterase

Pexel *Plasmodium* export element

PNP Purine nucleoside phosphorylase

PSSM Position specific scoring matrix

PV Parasitophorous vacuole

RBC Red blood cell

RSS Regular secondary structure

VTS Vacuolar transport signal

SG Structural Genomics

Chapter 1

Introduction

Genome sequencing projects have generated large amounts of sequence information. To gain a better understanding of biological systems on the molecular and physiological levels, structures and functions need to be assigned to all known and predicted proteins. Various experimental and computational techniques are available for the accomplishment of such tasks. Generating experimental data to provide evidence of protein structure and function is expensive, difficult and slow. Predictive computational methods are fast and applicable to whole proteomes, but are not as reliable as experimental results. However, predictions identify proteins of interest and determine their suitability for experimental studies (Liu *et al.*, 2004; Frishman, 2002). Furthermore, knowledge of structural features of proteins provides guidance for designing experiments (Herrera *et al.*, 2007).

Currently, there is a deficiency in protein structures compared to the amount of protein sequences available. Three dimensional (3D) protein structures determined by X-ray crystallography, Nuclear Magnetic Resonance (NMR) and Electron Microscopy are stored in the RCSB Protein Data Bank (PDB; Berman *et al.*, 2000). In contrast, protein sequence information is available from several publicly available databases. The SWISS-PROT (Bairoch *et al.*, 2004) component of the UniProt Knowledgebase (UniProtKB, Wu *et al.*, 2006) is the most frequently used database of manually annotated proteins (`http://www.expasy.org/sprot/`). Associated with SWISS-PROT, is a computer-annotated supplement database, UniProtKB/TrEMBL, which consists of all translated sequences from EMBL nucleotide sequence entries which have not yet been

Table 1.1: Comparison between the availability of protein sequence and 3D structure data.

Database	Protein information	May 2006	May 2007	Growth
PDB	Experimental 3D structures	33 982	39 920	5 938
SWISS-PROT	Manually annotated sequences	219 361	267 354	47 993
TrEMBL	Computer-annotated sequences	2 914 826	4 361 897	1 447 071

incorporated into SWISS-PROT (`http://www.expasy.org/sprot/`). The shortage in 3D structures for available protein sequences is summarized in Table 1.1. In addition to the difference in sequence-structure information, there is a large and increasing difference in the amount of protein sequences that are manually and computer-annotated. Consequently, structural and functional annotation of proteins is relying increasingly on computer-based predictive methods.

Methods used for protein structure prediction are primarily based on homology transfer. Structure is more conserved than sequence and therefore distantly related sequences often have the same or very similar structures (Chothia and Lesk, 1986). Computational methods for structure feature prediction make use of neural networks, statistical methods and physical properties of amino acid sequences. Typical computer-based methods for structural annotation include prediction of secondary structure, transmembrane helices, low complexity, disorder, coiled-coils, and 3D structure.

There is an increasing need for integrating different structural and functional annotations for three major reasons: Different databases cover different sets of proteins; prediction methods have different strengths and weaknesses and finally, biological conclusions about function and structure can be drawn more accurately considering as much information as possible about a certain sequence. Many meta-servers and integrated databases for protein structural and functional annotation have been generated in response to the need for integration of all the different types of information. Genome-scale annotation databases have been constructed by applying meta-servers to all the proteins encoded by the genome of a certain organism, collectively called a proteome. Proteome annotation with regard to structure and function is important for comparative studies (Liu and Rost,

2001) and for selecting sets of proteins of particular interest from an organism (Liu *et al.*, 2004). For many genome annotation databases, the human genome and various bacterial genomes have taken first priority for annotation (Carter *et al.*, 2003; Frishman *et al.*, 2003). More specifically, bacterial genomes are small and computationally less expensive to annotate than the larger eukaryotic genomes, and therefore are more readily annotated. Additionally, structural genome annotation of bacterial and eukaryotic pathogens, such as the malaria parasite, are especially important for the identification of new drug target candidates and the elucidation of molecular pathology within the organism. In the following section, the malaria parasite and its proteome will be discussed with regard to the importance of structural information on the proteins of this organism.

1.1. Malaria and the *Plasmodium falciparum* proteome

Malaria is a life threatening, infectious disease occurring in tropical and subtropical regions across the world. It infects approximately 300 million people per year and results in over one million deaths, predominantly in children from sub-Saharan Africa. Human malaria infections are caused by four protistan parasites: *Plasmodium falciparum, Plasmodium malariae*, *Plasmodium ovale* and *Plasmodium vivax*. *P. falciparum* causes the most lethal form of malaria and is transmitted to humans by female mosquitos of the genus *Anopheles*. Malaria parasites undergo different life stages in the human host; sporozoites first infect liver cells. Following infection, sporozoites develop into merozoites which penetrate red blood cells (RBC), where they reproduce asexually. The blood stage of the life cycle is responsible for all pathogenic effects, in addition to immune evasion and drug resistance (Miller *et al.*, 2002). One of the main problems contributing to the malaria epidemic, is the increasing resistance of *Plasmodium* to existing anti-malarial drugs.

Resistance to therapeutic drugs such as sulfadoxine and pyrimethamine have developed at large over the past two decades (Brooks *et al.*, 1994; Peterson *et al.*, 1988).

Following the arising resistance, there has been a pressing need to understand the mechanism of drug resistance and develop novel anti-malarial drugs. Protein structure has previously been used to elucidate the mechanism of resistance in *P. falciparum* (de Beer *et al.*, 2006; Yuthavong *et al.*, 2005). Furthermore, inhibitors can be designed from structure (Sarma *et al.*, 2003; Velanker *et al.*, 1997). As resistance to existing drugs is a globally occurring phenomenon, new information regarding the structure and function of the proteins in the *P. falciparum* genome is of importance.

With the recent completion of the *P. falciparum* genome sequencing project (Gardner *et al.*, 2002), many proteins await functional and structural characterization. However, various features of the parasite genome and proteome complicate functional and structural characterization studies. The *P. falciparum* genome is exceptionally AT-rich (Gardner *et al.*, 2002) giving rise to an abundance of low complexity regions in the protein products. Low complexity is a characteristic of parasite-specific inserts. In the context of a multiple alignment of a specific protein, inserts refer to regions in the parasite protein not present in homologous proteins from other species, separating the aligned regions. They are often Asn-rich and contain polar amino acids (Rozmajzl *et al.*, 2001). Inserts also cause *P. falciparum* proteins to be considerably longer than their homologs in other species. Consequences of these characteristics include difficulties with cloning, purification, expression and crystallization of proteins and thus these difficulties are reflected in the low number of 3D structures that have been experimentally determined for *P. falciparum* proteins.

A search of the PDB using "falciparum" as a keyword rendered 210 structures and removal of similar sequences with more than 90% sequence identity, rendered 103 structures (`http://www.rcsb.org/pdb/`). The Gene Ontology (GO) database is a functional classification of proteins on cellular, molecular and biological levels, making use of a controlled vocabulary (Ashburner *et al.*, 2000). GO terms have been assigned manually to 40% of all *P. falciparum* gene products (Gardner *et al.*, 2002). Almost 60% percent of

the proteins do not have sufficient similarity to known proteins and therefore no function can be assigned to them.

In summary, 4% of the *P. falciparum* proteins have experimental 3D structures assigned, and 60% of the proteome is described as hypothetical. Furthermore, the amount of redundant *P. falciparum* proteins in the PDB is significant. Therefore, the identification and prioritization for experimental and computational structural studies will speed up protein characterization and prevent repetition of expensive experiments. Prioritising proteins based on uniqueness and biological importance is dependent on high-throughput computational structural and functional feature annotation.

1.2. Computational methods for protein structure determination

In order to gain concise structural information from protein sequences, different approaches for structure annotation were integrated. In the following sections, bioinformatics tools and methods available for genome-scale structural annotation using sequence information alone will be reviewed. These methods were integrated to structurally annotate the *P. falciparum* proteome (Chapter 2). Subsequently, the annotations were utilized to perform the putative identification of protein targets for structural experimental and computational studies (Chapter 3). With the exception of fold recognition, 3D structure prediction will not be extensively addressed. Search methods will be mentioned and briefly explained where appropriate, but an in-depth review of search methods is beyond the scope of this discussion.

1.2.1. Structure template recognition

Structure templates are protein sequences with known 3D structures which can be used to predict the structure of a similar sequence. Templates are found in structural databases such as the PDB, SCOP (Murzin *et al.*, 1995) and CATH (Pearl *et al.*, 2005). The PDB contains structures of whole protein sequences in contrast to SCOP and CATH which contain structural domains of proteins. Templates can be searched for by several

methods which are classified into four types: Sequence-sequence comparison methods, profile-sequence comparison methods, profile-profile comparison methods and sequence-structure alignments or threading. The first three comparison methods are collectively referred to as sequence-based comparison methods and can be applied to search the PDB, SCOP or CATH.

Sequence-based comparison methods

Sequence-sequence comparison methods are computationally the least expensive. The most well known example of this type of method, for structure template identification, is a BLASTP (Altschul *et al.*, 1990) search against the PDB. From multiple sequence alignments, positions of high conservation and high variability can be detected. To find sequences that are related to the family of sequences in the multiple alignment, positions are given weights. If a target sequence aligns at a certain position to an alignment position with a high weight, the score will increase. However, if the target sequence misaligns at a position with a lower weight, the score will not be affected as much. Such a weighted alignment is called a profile.

Two types of profiles can be generated from multiple alignments. They include Position Specific Scoring Matrices (PSSMs; Dayhoff *et al.*, 1974) and Hidden Markov Models (HMMs; Krogh *et al.*, 1994). PSSMs and HMMs are statistical models describing amino acid frequencies at each position in a multiple alignment and are used to search for homologs. A profile generally returns fewer false positives because it is matched to entire domains, although it is less sensitive at the most conserved part of the sequences. A profile-sequence comparison method includes Position Specific Iterated BLAST (PSI-BLAST; Altschul *et al.* 1997) which uses the outputs of a BLAST search to construct a sequence-specific PSSM. The PSSM is used as input for a next BLAST search and this process is repeated for a user-specified amount of times. HMMER (Eddy, 1996) and SAM-T99 (Karplus *et al.*, 1998) are two more examples of profile-sequence comparison methods. These packages can be used to construct HMMs from multiple

alignments and search databases of HMMs. SAM-T99 also has an accurate multiple alignment program.

In separate studies, the performance of HMM packages were compared with PSI-BLAST, WU-BLAST, and NCBI-BLAST (Madera and Gough, 2002; Edgar and Sjolander, 2004). The profile-sequence comparison methods outperformed the sequence-sequence methods by more or less 40% in both cases (Figure 1.1). SAM-T99 is considered as the state of the art method for HMM building, and this method also outperforms PSI-BLAST as illustrated by Edgar & Sjolander and Madera & Gough. The SAM-T99 method is especially good in finding and aligning templates for comparative modeling studies (Ginalski *et al.*, 2005).

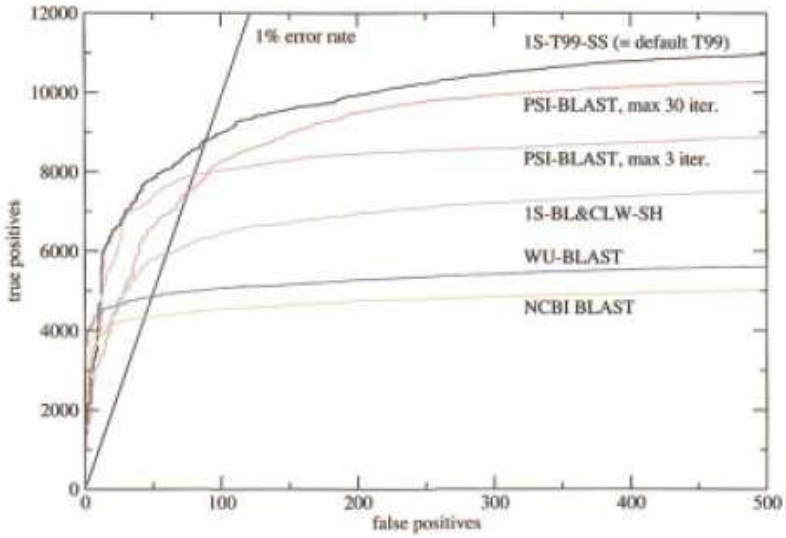


Figure 1.1: Sensitivity plots for the SCOP (1.5) all-against-all. 2 873 sequences were used that had less than 40% sequence identity, with a total of 36 612 possible true pairwise relationships. 1S is a single sequence seed, BL&CLW-SH is the result of a simple WU-BLAST search, aligned with ClustalW, T99 is the alignment procedure used by SAMT99, SS is the default SAMT99 (Madera and Gough, 2002).

A protein domain is defined as a stretch of sequence that folds to form an independent functional structural unit (Murzin *et al.*, 1995). It is therefore useful to search for structural domains within a query protein sequence if the sequence could not be matched to any protein in the PDB. Structural domain databases include SCOP and CATH. HMM

libraries representing the domains in SCOP and CATH have been constructed and include Superfamily (Gough *et al.*, 2001) and CATH-HMM.

SCOP (`http://www.biochem.ucl.ac.uk/cgi-bin/cath/CathServer.pl`) is a classification system for protein structures and was generated from structural alignments of proteins. The structural alignments are constructed by aligning pairs of residues that occupy the same geometric positions in two protein structures. Hierarchical clusterings resulted in four domain classifications in SCOP:

1. Homologous domains are grouped to form a family.
2. Families that share a similar structure and function, but not significant sequence similarity, form a superfamily.
3. Superfamilies that share a common folding topology are in turn grouped to form a fold.
4. Finally, fold groups fall into one of the general classes which include: all-alpha (α), all-beta (β), alpha and beta (α+β), alpha and beta (α/β), multi-domain proteins, membrane and cell surface proteins and miscellaneous small proteins.

As profile-sequence comparison methods are more sensitive than sequence-sequence comparison methods, HMMs representing SCOP superfamilies were constructed and reposited in the Superfamily database (`http://supfam.org`). The aim of the Superfamily database is to identify domains within all known protein sequences that belong to superfamilies of known structure. The library of HMMs has been applied to all fully sequenced genomes as well as all sequences in UniProt. All results are available for viewing and downloading. Tools to search the HMM library are also available for download. Superfamily gives a summary for each genome of over- and under-represented superfamilies. Unusual combinations of domains in a certain genome can also be viewed. The current Superfamily procedure compares a query sequence against the profile HMMs in the database. A program is in the process of development for comparison of two profile HMMs for detection of more remote homologs (Madera *et al.*, 2004). Another SCOP superfamily-related profile

database is SUPFAM (Pandit *et al.*, 2002).

SUPFAM (`http://pauling.mbu.iisc.ernet.in/~supfam`) describes relationships between Pfam domain families and SCOP superfamilies. A SCOP-derived database, PALI (Gowri *et al.*, 2003) contains families of a homolog with a known 3D structure and structural alignments of these families. A PSI-BLAST PSSM was constructed for every Pfam family and every PALI family. The Pfam PSSMs and PALI PSSMs were then matched by Reversed Position Specific BLAST (RPS-BLAST) searches. RPS-BLAST searches a sequence against a database of profiles (the opposite of PSI-BLAST which searches a profile against a database of sequences).

CATH is a database of protein domain structures and classifications similar to SCOP classifications. The database currently contains 43 229 domains classified into 1 467 superfamilies and 5 107 sequence families. Structural families are extended to sequence relatives from GenBank and extended CATH contains 616 470 domain sequences classified into 23 876 sequence families. The Dictionary of Homologous Superfamilies (DHS) contains sequence, structure and functional information for each superfamily in CATH. HMMs were built for representative sequences from each sequence family. More remote homologs can be identified by scanning against the CATH-HMM library.

Threading

Threading is a fold recognition method which uses experimental structural information and subsequent energy calculations for predicting the 3D structure of a protein sequence. Instead of comparing two sequences, or an alignment derived-profile with a sequence, it tries to fit an amino acid sequence onto the backbones of a set of known protein structures. For each sequence-structure alignment, a goodness-of-fit score is calculated based on the forcefields of the observed protein structures. In effect, any query sequence is compared to the fixed set of protein structures known as a fold library. Consequently, threading relies on the assumption that there is a finite number of possible protein folds in the

universe (Chothia, 1992).

The word 'threading' was coined by T.D. Jones, who also developed one of the first threading algorithms, Threader (Jones *et al.*, 1992). Threader has continuously been improved since the first version and now incorporates secondary structure predictions of the queried sequences to improve alignments (Jones *et al.*, 1999). The most recent version is Threader 3, which also includes the Threading Expert (Texp). Texp is used for the automatic interpretation of Threader results and identification of the most suitable template for the unknown sequence. Threader 3 is freely available for download (`http://bioinf.cs.ucl.ac.uk/threader/threader.html`).

Threading can successfully pair an unknown sequence with a fold (Jones *et al.*, 1995; Shortle, 1997), however there are a few disadvantages which include: Inaccurate sequence-structure alignments (Lemer *et al.*, 1995); the manual interpretation of threading results is crucial and finally, the limited potential for application to whole proteomes as a result of the computational time that is required (Jones, 1999*a*). Threading is therefore appropriate for relatively short sequences and can be successfully applied only when results will be interpreted manually. Since a whole query sequence will be aligned to a structural fold even though the sequence might consist of more than one structural fold, domain identification is very important for the success of threading (Jones and Hadley, 2000; Marsden *et al.*, 2002). Meta-predictors generally perform better than single threading methods without human intervention (Ginalski and Rychlewski, 2003), but are difficult to apply on a genome-scale and also incorporate sequence-based comparison methods.

1.2.2. Secondary structure prediction

Secondary structure prediction plays a big role in protein tertiary structure prediction (Boscott *et al.*, 1993), increasing alignment accuracy (Jennings *et al.*, 2001) and protein function prediction (Jensen *et al.*, 2003). Secondary structure prediction methods take

one or more protein sequences as input and predict for each amino acid one of three secondary structure states. The three states include helix (H), strand (E) or coil (C). A myriad of programs exist for the prediction of protein secondary structure making use of HMMs, Artificial Neural Networks (ANNs), Support Vector Machines (SVMs) as well as some older nearest-neighbour approaches. One of most consistent best-performing methods is the PSIPRED method (Jones, 1999*b*). This neural network-based method uses a PSI-BLAST PSSM as input. The method's accuracy was increased by taking a consensus prediction from four independent neural networks and the Q3 score currently averages 79.6% on the EVA (evaluation; Grana *et al.*, 2005) set. EVA is a web service which routinely evaluate structure prediction algorithms as new structures become available in the PDB. A hybrid method, HYPROSP II (Lin *et al.*, 2005) makes use of three components for prediction: a combination of PSIPRED and PROSP, a sequence-structure knowledge base and a voting system. HYPROSP II achieves an average Q3 accuracy of 80.7% on the EVA set. Other methods that perform well are PROFsec, SAM-T99sec and SSpro4 1.2. Performance results for PROFsec, SAM-T99 and PSIPRED for the last year are available from the EVA web site (`http://cubic/bioc.columbia.edu/eva`) and are compared in Figure 1.2.

1.2.3. Transmembrane helix prediction

Membrane proteins are estimated to form 25% of all proteins in a proteome (Krogh *et al.*, 2001). However, there are only 638 alpha helical transmembrane protein structures in the PDB as detected by the TMDET algorithm (Tusnady *et al.*, 2004; `http://pdbtm.enzim.hu/`) as a result of the difficulties with solubilization, purification and crystallization of membrane proteins. Nevertheless, membrane protein structures are of utmost importance as they are responsible for activities such as: Transport of ions and metabolites; energy generation through electron flow-coupled ATP synthesis; and signal transduction of neurotransmitters, growth factors and hormones across the membrane.

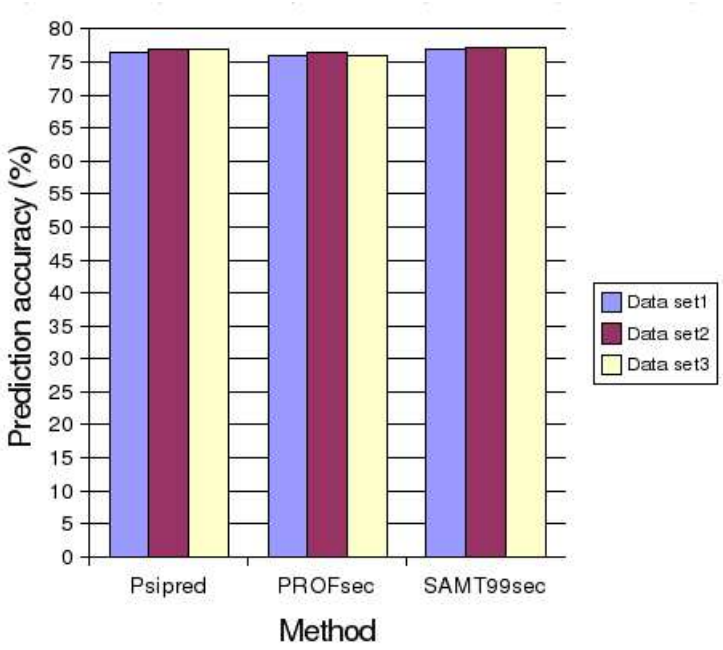


Figure 1.2: Comparison of the accuracies obtained on three EVA data sets by three secondary structure prediction methods (`http://cubic/bioc.columbia.edu/eva`). These methods obtained consistent accuracies above 75% for all the datasets.

The abundance, accessibility and functional importance of membrane proteins make them crucial drug targets.

Methods for prediction of transmembrane helices are usually based on two signals. The first signal is hydrophobicity. Regions that span the membrane contain a high propensity of hydrophobic amino acids, due to the hydrophobic environment in lipid membranes. The second signal is the abundance of positively charged residues in the loops that are located on the cytoplasmic side of the membrane. This "positive inside rule" is used for predicting the orientation of a transmembrane helix (von Heijne, 1989). In addition, helical membrane proteins follow a pattern in which the cytoplasmic and non-cytoplasmic loops are alternating.

Algorithms for transmembrane helix prediction should ideally be able to incorporate all of the above mentioned signals. HMMs are well suited for transmembrane prediction because they enable incorporation of a large range of variables into one model. Therefore, many methods for prediction use the HMM approach (Krogh *et al.*, 2001; Kall *et al.*, 2004;

Kahsay *et al.*, 2005). Another successful approach is the utilization of neural networks (Rost *et al.*, 1995). TMHMM (Krogh *et al.*, 2001) is a widely used HMM-based prediction method and will be discussed briefly as an example.

TMHMM predicts a transmembrane helix by finding the most probable topology given the HMM. HMM parameters include the probabilities of each of the 20 amino acids in the different states as well as the lengths. These parameters are estimated from 160 transmembrane proteins of known structure. TMHMM falsely predicts approximately 20% of signal peptides as transmembrane helices for eukaryotes, and 60% for gram positive prokaryotes (Krogh *et al.*, 2001). Table 1.2 compares the prediction accuracy of TMHMM2 to other transmembrane helix predictors. The TMHMM prediction server can be accessed at `http://www.cbs.dtu.dk/services/TMHMM/`.

Table 1.2: Comparison of accuracies for several TM prediction methods (extracted from Cuthbertson *et al.*, 2005). Np is the amount of predicted transmembrane helices, Nc is the amount of correctly predicted TM helices out of 268 observed, Qp is the per segment accuracy, and Q3 is the overall per residue accuracy for all residues.

Prediction Method	**Np**	**Nc**	**Qp (%)**	**Q3 (%)**
HMMTOP2	270	254	94.4	82.2
SPLIT4	254	250	95.8	85.2
TMHMM2	246	239	93.1	83.3
DAS	297	260	92.2	80.7
TMAP	249	240	92.9	81.9

1.2.4. Coiled-coils

A coiled-coil is a bundle of α-helices which are wound around each other to form a superhelix. The defining feature of a coiled-coil is an amino acid heptad repeat in which the first and fourth residues are hydrophobic and the fifth and seventh residues are charged or polar. Such a repeat gives rise to the 'knobs-in-holes' packing of α-helices as described by Francis Crick (1953). Coiled-coils are crucial for cellular function as they are responsible for holding together molecules, subcellular components and tissues. Many diseases such as progeria, cancer and neurodegenerative diseases, are associated

with mutations in coiled-coils (Chigira *et al.*, 2003; McClatchey, 2003; Puls *et al.*, 2003). Prediction of coiled-coils therefore facilitates a better understanding of protein structure and function. Coiled-coils are of further interest as heptad repeats cause regions of low complexity. It is therefore useful to mask coiled-coil regions when searching for homologs (Rose *et al.*, 2005).

Methods predicting coiled-coils search for a motif that corresponds with the heptad repeat. For most methods, at least four successive heptad repeats are necessary to predict a coiled-coil structure. A large number of consecutive heptad repeats increases the confidence of coiled-coil predictions. COILS (Lupas *et al.*, 1991) predicts coiled-coil structures in a protein sequence using similarity searches between a query sequence and a database of coiled-coil proteins. COILS can be accessed at `http://www.ch.embnet.org/software/COILS_form.html`. Paircoil2 (McDonnell *et al.*, 2006) is an improved version of Paircoil, using pairwise residues probabilities for predicting coiled-coils. It has been reported that Paircoil2 has 98% sensitivity and 97% specificity. Multicoil is an extension of Paircoil to three stranded coiled-coils (Wolf *et al.*, 1997) and can be accessed at `http://theory.lcs.mit.edu/multicoil`.

1.2.5. Low complexity regions

A globular protein has a spherical shape with hydrophobic residues buried in its core and hydrophilic residues exposed to the outside. Some proteins are primarily globular, but contain regions of non-globularity or low complexity. Not much is known about the function, structure and interactions of these regions. SEG (Wootton, 1994) is an algorithm to predict regions of low complexity in protein sequences. For many search methods it is useful to mask low complexity regions as these regions are more likely to produce random false positives hits. In addition, low complexity prediction helps identify domain boundaries and globular regions in the protein. Low complexity regions are abundant in *P. falciparum* proteins and are therefore an important feature to annotate in the *P.*

falciparum proteome. Regions which do not form regular secondary structure are related to low complexity although these regions can become more structured upon binding to other molecules.

1.2.6. Unstructured regions

Unstructured or disordered regions in proteins are defined as regions lacking any regular secondary structure (RSS). RSS includes helices, strands, membrane helices, signal peptides and coiled-coils. Unstructured regions may play important functional roles in facilitating protein folding, scavenging and post-translational modification (Romero *et al.*, 1998; Tompa, 2005). A well-studied example of the role of disorder in protein-protein interactions, includes the myosin-actin complex (Rayment *et al.*, 1993). The correlation between interactions and disorder is logical since interacting regions will undergo change in conformation more easily when these parts are more flexible than the rest of the protein. Eukaryotes contain more proteins with long disordered regions than prokaryotes and archae. These proteins or regions in proteins are difficult to crystallize and are often invisible in electron density maps, and it is therefore of interest to know which proteins may contain unstructured regions, before X-ray crystallization studies are attempted.

Romero *et al.,* (2001) characterized intrinsically disordered regions as having low sequence complexity and developed a neural network based method, PONDR (Predictor Of Natural Disordered Regions; Romero *et al.*, 2004) for predicting disorder. Currently, PONDR is only available commercially (`http://www.pondr.com/`).

Liu and Rost (2002) investigated unstructured regions in the PDB and developed a method to predict regions of no regular secondary structure (NORS) from protein sequences. The method is based on a whole range of preliminary predictions such as secondary structure, transmembrane helices, coiled-coils, signal peptides and solvent accessibility. Predictions are made by setting criteria with respect to the above mentioned

predictions for a sliding window of 70 residues.

The DisProt/VSL2 package utilizes machine learning approaches as well as evolutionary information for the purpose of predicting regions of intrinsic disorder (Peng *et al.*, 2006). The average prediction accuracy for long and short disordered regions is 81%. The prediction accuracy depends on the following optional inputs given to the program: A PSI-BLAST PSSM file, a PSIPred secondary structure prediction file and a PHD secondary structure prediction file. DisProt/VSL2 is freely available for non-commercial use and can be downloaded at `http://www.ist.temple.edu/disprot/predictorVSL2.php`. Other programs for predicting intrinsic disorder includes Disopred (Ward *et al.*, 2004*a*), DisEMBL (Iakoucheva and Dunker, 2003) and IUPred (Dosztanyi *et al.*, 2005). All of these methods are freely available for academic use.

1.2.7. Family recognition and domain identification

Proteins within a family often share the same or a similar function. Assigning an unknown protein to a family of proteins can therefore indicate the protein's function. Family assignment takes place by recognition of a functional domain within a protein query sequence. As mentioned earlier, function transfer through homology is often incorrect and should be verified by alternative methods. However, domain identification is useful for structure prediction methods such as threading and homology modeling. Using the domain, rather than the whole protein sequence as input for these methods, results in more accurate template identification. This section will briefly describe the most well known protein family databases.

Pfam

Pfam is a database containing a collection of multiple alignments, profile-HMMs and annotations for each alignment (Finn *et al.*, 2006). Release 21.0 of Pfam contains 8 957 manually curated families (`http://pfam.wustl.edu`) and is based on SWISS-PROT and

SP-TrEMBL protein databases. The aim of Pfam is to assign domains of known function to novel protein sequences. For each family in Pfam-A there are a FULL alignment, a SEED alignment, a ls-HMM, a fs-HMM and annotations. SEED alignments are constructed by ClustalW and T-Coffee, using protein sequences that are most representative of the family. FULL alignments contain all known examples of a particular family. The largest family in 2002 was the HIV GP120 glycoprotein containing more than 27 000 sequences in the full alignment (Bateman and Haft, 2002). For each family there are two HMMs. The first HMM is generated in the ls or global mode of HMMER. Global mode means that the whole HMM must be aligned to the sequences but is aligned locally with respect to the sequence. The other HMM is in fs or local mode. Local mode means that only a part of the HMM is aligned. Thus, an fs-HMM can be used to search partial sequences like expressed sequence tags (ESTs). The most sensitive search of Pfam would include ls-mode and fs-mode searches at the expense of doubling the amount of profile-HMMs to search.

Pfam-B is an automatically derived collection of protein families, generated by taking all families in ProDom (Servant *et al.*, 2002) and removing those that are already in Pfam-A. Pfam-B families do not contain profile-HMMs, annotation or full alignments. Thus, Pfam-B covers more families but is not curated and of much lower quality than Pfam-A.

SMART

SMART (Simple Modular Architecture Research Tool; Letunic *et al.*, 2006) is a similar database to Pfam but much smaller. SMART contains HMMs of domain alignments which were manually derived. The database is used to identify domains in protein sequences. Additional information includes enzyme active sites and putative protein-protein interactions. SMART contains more than 685 entries, 500 of these representing domains contained in extracellular eukaryotic signaling and chromatin-associated proteins

(`http://smart.embl.de/`).

TIGRFAMs

TIGRFAMs (Haft *et al.*, 2003) is a database containing manually curated protein families, HMMs, multiple sequence alignments, commentary and Gene Ontology assignments. TIGRFAMs was designed from the start to be complimentary to Pfam, whose models are representative of a domain and not the whole sequence. The TIGRFAMs database currently contains more than 1 600 protein families and is available for downloading (`http://www.tigr.org/TIGRFAMs`).

SYSTERS

The SYSTERS database (Meinel *et al.*, 2005) is a fully automated classification of proteins into a family and superfamily. It is derived from rigorous all-against-all Smith-Waterman searches. The resulting pairs of sequences are clustered in a refined two-step approach. The SYSTERS clustering method makes use of distance graphs to cluster proteins first into superfamilies and then into families. The web server, `http://systers.molgen.mpg.de` provides access to 158 153 SYSTERS protein families. The SYSTERS data comprise several protein sequence databases derived from completely sequenced organisms, including ENSEMBL, TAIR, SGD, GeneDB and SWISS-PROT/TrEMBL.

ProDom

ProDom contains all protein domain families automatically generated from the SWISS-PROT and TrEMBL sequence databases (Servant *et al.*, 2002). The November 2005 release of ProDom was built from 1 067 651 non-fragmentary protein sequences. The non-fragmentary protein sequences were clustered into 736 449 domain families, with 275 561 families containing at least two sequences. Domain families are generated by a clustering program known as MKDOM2 (Gouzy *et al.*, 1999). This program is based on

the assumption that the shortest amino acid sequence corresponds to a single domain protein not including fragmentary sequences. ProDom-CG is an extraction of the standard ProDom release and contains domains that belong to complete genomes. ProDom-CG 2001 contains 171 complete genomes, including *P. falciparum*. A comparison of the coverage of different family databases can be seen in Figure 1.3. The difference in coverage between partially automated/partially manual and fully automated annotation databases is clearly illustrated. Since motif databases are also used for family identification, the coverage comparison in Figure 1.3 includes motif databases.

CDD

The conserved domain database (CDD) is a collection of conserved domains maintained by NCBI Entrez (Marchler-Bauer *et al.*, 2005). CD-search is a tool to search a sequence against CDD and can be accessed at `http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi`. The Conserved Domain Database consists of domain models from Pfam, Clusters of Orthologous Groups (COG, Tatusov *et al.*, 2003), SMART and several hundred NCBI-curated domain models. In order to reduce some of the redundancy between these groups of domains, overlapping hits of the models in the protein database are clustered. Members that do not add significantly to the cluster's coverage are subsequently removed from the CDD collection. CDD contains 5 252 of 7 255 Pfam version 11.0 models, 575 of 663 SMART version 4.0 models, and 4 101 of 4 873 COG models.

InterPro

InterPro (Mulder *et al.*, 2005) is an integrated database containing family, domain and functional site information from Pfam, SMART, TIGRFAMs, ProDom, Panther (Thomas *et al.*, 2003), Gene3D (Yeats *et al.*, 2006), Superfamily, PRINTS (Attwood *et al.*, 2003), PROSITE (Bairoch, 1991), and PIRSF (Wu *et al.*, 2004*b*). The family entries of InterPro are compared to the single database entries in Figure 1.3. InterPro will be discussed in

Chapter 2 as an example of an integrated database.

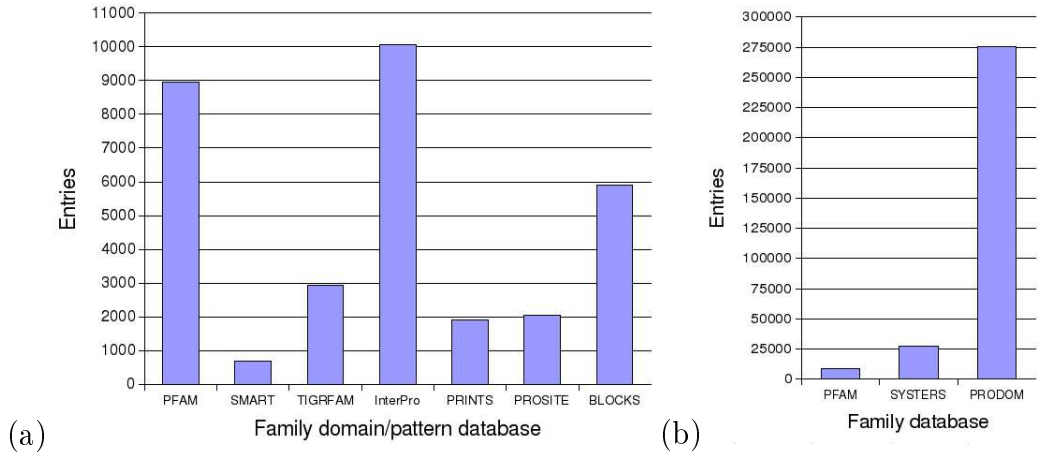


Figure 1.3: Comparison of family databases

(a) Comparison of family coverage by different databases, reflecting the amount of entries representing protein families (b) comparison of automatic family domain databases with manually annotated Pfam.

1.2.8. Motif and pattern searching

Motif or pattern searches do not rely on whole sequence similarity to other sequences, but rather the conservation of a few critical residues, called a pattern, signature or a motif. Patterns that are functionally important tend to be more conserved through evolution than the rest of the protein sequence. Motifs or patterns are used to identify active sites and protein families. Additionally, motifs indicate positions of post-translational modification (e.g. glycosylation sites), structural signals (e.g. N and C caps of α-helices), or signal sequences. Motifs can be recognized by making use of regular expressions or PSSMs. From a set of similar sequences containing a particular motif, regular expressions can be derived. Regular expressions are used to describe the permitted amino acids in each of the motif positions. Regular expressions allow for different amino acids in one motif position and are therefore flexible. However, regular expressions do not permit gap insertion. Motif databases often contain multiple alignments of motifs and corresponding PSI-BLAST constructed PSSMs of the alignments. These PSSMs can then be searched

by a RPS-BLAST search. Function is often associated with multiple motifs which are separated in the sequence. Single motif databases can therefore falsely assign function to proteins. In response to this weakness, multiple-motif databases have been generated. Examples of motif databases include PROSITE, PRINTS and Blocks (Henikoff *et al.*, 2000).

PROSITE

PROSITE is a database containing patterns of biologically important amino acid sequences. The patterns in PROSITE are described as regular expressions. Profiles of multiple alignments of domains have also been added to the PROSITE database. Search tools for PROSITE include ScanProsite and Patmatmotifs from EMBOSS. PROSITE Release 20.13 (May 2007) contains 1 484 documentations describing 1 319 patterns and 734 profiles. ProRule is a set of 746 rules providing information about functionally and structurally crucial amino acids. ProRule is used in addition to patterns and profiles to discriminate more effectively between true and false matches (`http://www.expasy.org/prosite`).

PRINTS

PRINTS is a database consisting of fingerprints and their annotations. A fingerprint is a group of motifs that together constitute characteristics of a certain protein family. The PRINTS database is manually maintained and therefore rather small, containing 1 904 fingerprints derived from 11 451 individual motifs. SWISS-PROT/TrEMBL is used as a source for protein sequences. To increase the size of the database, an automatic supplement PrePRINTS has been developed. PrePRINTS is generated by a pipeline which uses protein family clusters from ProDom as input. Motifs are detected using a suite of programs which are subsequently used to search the UniProt database in an iterative way. PrePRINTS contains 250 entries that await manual refinement. PRINTS is available freely at `http://umber.embnet.org/dbbrowser/PRINTS` and can be down-

loaded *via* FTP (`ftp://ftp.bioinf.man.ac.uk/pub/prints`).

Blocks

The Blocks database (`http://blocks.fhcrc.org/`) consists of blocks of ungapped multiple alignments corresponding to the most conserved regions of proteins (Henikoff *et al.*, 2000). Similar sequences contain the same set of non-overlapping blocks, generated by the PROTOMAT system (Henikoff and Henikoff, 1991). New blocks are only added when no hits are obtained in searching the new block against the Blocks database. The Blocks database (release 14.3) contains 29 068 blocks forming 5 900 different groups. Blocks can be searched with BLIMPS and IMPALA (Schaffer *et al.*, 1999) to compare a DNA or protein sequences with the blocks in the database. Similarly to RPS-BLAST, IMPALA searches a database of PSI-BLAST PSSMs. However, IMPALA performs a SMITH-WATERMAN calculation between the query and each profile rather than finding word-hits that could be extended (Schaffer *et al.*, 1999). These two search methods yield the same true positives, but different false positives. It is therefore useful to compare results for the two programs (Henikoff *et al.*, 2000).

Signal Peptide prediction

A signal peptide at the N-terminal of a protein, targets a sequence to a specific subcellular location. The subcellular location is indicative of the functional role of the protein. The oldest methods for prediction of signal peptides involved weight matrices, such as SigCleave (von Heijne, 1986). SignalP (Bendtsen *et al.*, 2004) is one of the most popular methods to predict signal peptides. SignalP uses combined ANN and HMM approaches to predict the cleavage sites of signal peptidase I. A separate program has been developed for signal peptidase II cleavage site prediction in lipoproteins (Juncker *et al.*, 2003).

Predictions with specific relevance to *P. falciparum* proteins have been made based on sequence motifs. A set of proteins are exported from the parasitophorous vacuole

(PV) of the parasite into the RBC. These proteins are responsible for remodeling and modifications of the RBC which ensure parasite survival (Haldar *et al.*, 2002) and invoke pathogenesis (Miller *et al.*, 2002). Exported proteins are therefore good candidates for drug targets. Two research groups have independently predicted all the proteins in the *P. falciparum* proteome which are exported from the parasite to the human red blood cell (RBC), based on two motifs (Hiller *et al.*, 2004; Marti *et al.*, 2004). The N-terminal signal sequence, referred to as the SS signal, facilitates translocation of proteins into the endoplasmic reticulum (Wickham *et al.*, 2001). Subsequently, the exported proteins are transported across the parasite membrane into the PV. Marti *et al.* identified a pentameric motif, the *Plasmodium* export element (Pexel), and Hiller *et al.* identified an 11-amino acid motif which they named the vacuolar transport signal (VTS). These two motifs are commonly called the Pexel/VTS motif and are necessary for the export of proteins across the PV membrane into the RBC. Consequently, a proteome search of the SS signal together with either the Pexel or the VTS motifs, lead to the predictions of the *P. falciparum* secretome.

Many of the proteins predicted to be exported, are hypothetical proteins and lack any description regarding function. However, knowledge that the proteins are exported, leads to the assumption that these proteins are involved in parasite survival and pathogenesis. Therefore, these proteins are responsible for drug resistance, cytoadherence, rosetting, antigenic variation, or induced lipid and protein trafficking in the RBC. Together with additional protein sequence analysis, possible functions for the protein can be narrowed down. Signal peptide prediction and motif recognition are valuable for functional annotation and putative drug target identification.

1.3. Conclusions

Experimental methods for structure determination cannot keep up with the pace at which new protein sequences become available. Various computational approaches for

annotating structure and structural features to whole proteomes are available. Structural and functional characterization of proteins is crucial for identifying novel drug targets and for structure-based inhibitor development. As resistance towards anti-malarial drugs is increasing and not many proteins have been experimentally characterized, computational proteome annotation of structural and functional features is necessary. *P. falciparum* structures in the PDB are redundant and targets for structural genomics need to be prioritized to prevent experiments from being repeated. One way to identify targets for structural studies is by proteome annotation of structural features, such as PDB coverage.

Computational approaches for structural annotation include: 3D template recognition, prediction of secondary structure, disorder, transmembrane domains, signal peptides, coiled-coils, functional motifs, low complexity regions and protein family or domain identification. For high-throughput annotation, stand-alone versions of prediction programs must be available. Many prediction programs are both performing well compared to similar programs in the field and freely available.

Several conclusions regarding specific computational methods and databases suitable for high-throughput annotation can be made. When searching for structural templates, profile-sequence comparison methods are more sensitive than simple sequence-sequence comparison methods, although the former is computationally more expensive. For example, PSI-BLAST and HMM-based searches are more sensitive than a BLAST search. If no homologs are found by sequence-based searches, threading should be attempted. However, threading requires substantial computational time and manual interpretation. Therefore, only sequences below a certain cut-off length, depending on the type of CPU, should be threaded. Alternatively, long sequences must be divided into domains, and these should be threaded separately.

Protein family databases are constructed either manually, automatically or by a combination of the two. While manually curated databases are more reliable than automatic

databases, the latter offers substantially more coverage than the former. Proteins should ideally be annotated by a combination of manual and automatic databases. In addition to functional annotation, protein family databases facilitate the identification of domains within a protein.

The predicted secondary structure improves the accuracy of 3D homology modeling and threading. Low-complexity, coiled-coil, disorder and transmembrane regions make a protein unfavorable for experimental structure determination. Therefore, knowledge of these features in proteins facilitates target selection for experiments. Prioritizing targets for structural determination, based on sequence uniqueness and functional importance, will increase the amount of protein sequences which can accurately be modeled by homology. Experimental targets should ideally have very low similarity to sequences in the PDB, as experimentally determined structures in the PDB are redundant.

Predicting sequence features in a proteome can indicate functional importance and novel drug targets as illustrated by the Pexel/VTS motif for export prediction in *P.falcipa-rum*. The exported proteins are involved in parasite survival and disease symptoms. Therefore, predicting all the proteins which are exported, facilitates novel drug target identification.

1.4. Problem statement

Resistance in malaria proteins continues to develop and new strategies for fighting the parasite have to be found. Protein structure plays a crucial role in elucidating mechanisms of parasite functioning and resistance. Protein structure furthermore aids the determination of protein function, which can together with the structure be used to identify novel drug targets in the parasite. Because of the abundance of low complexity, inserts and unstructured regions, *P. falciparum* proteins are problematic to crystallize. The lack of solved structures in the malaria parasite together with the redundancy of malaria protein structures in the PDB, necessitate the identification of suitable targets

for structure determination with novel structures. The identification of such targets can be accomplished by annotating structural features to the *P. falciparum* proteome.

1.5. Aims

In order to investigate protein structural features in the *P. falciparum* proteome the following aims were set:

1. Construction of an automated structural annotation pipeline on a Linux cluster (Chapter 2).
2. The use of the structural annotation pipeline to annotate the *P. falciparum, P. vivax* and *P. yoelii* proteomes (Chapter 2).
3. Representing structural features for each protein graphically and making the results available through a web-based interface (Chapter 2).
4. Putatively identifying suitable proteins for further in-depth experimental and computational three dimensional structural studies (Chapter 3).
5. Performing statistical analysis and comparison of structural features between the *P. falciparum, P. vivax* and *P. yoelii* proteomes (Chapter 3).

These aims will identify and prioritize functionally important targets for structural determination by experimental and computational methods. In turn, prioritization will increase the variety of known protein structures and facilitate homology modeling of similar proteins in *P. falciparum*.

Chapter 2

Structural annotation of *Plasmodium* proteomes

2.1. Introduction

Biological data integration is necessary for efficient data mining. Many resources for protein information exist. Protein sequence data is obtained from UniProtKB and the Protein Information Resource Sequence Database (PIR-SDB; Wu *et al.*, 2002). Protein sequence data is classified into superfamilies, families and subfamilies by different databases. Various resources for protein classifications focus on different sets of proteins and make use of different representative models (e.g. PSSMs and HMMs). Therefore, it is more informative but cumbersome to query all of the major protein databases for characterization of unknown protein sequences. For these reasons, protein databanks have collaborated to integrate their data into a single information resource. In addition to protein databases, numerous tools are available for predicting structural and functional features in proteins. The integration of protein data increases the potential for high-throughput data mining and provides an effective overview of the functional and structural features in proteins.

InterPro contains integrated protein information from PROSITE, PRINTS, ProDom, Pfam, SMART, TIGRFAMs, PIRSF, Superfamily, Gene3D and PANTHER. Models from these databases representing the same protein family have been assigned to a single InterPro entry. The resultant InterPro content covers 78% of all the proteins in UniProtKB. In

addition to protein domain family and superfamily information, InterPro offers information on protein structures modeled by MODBASE and SWISS-MODEL. Search methods for the component databases have been integrated into a single search tool known as InterProScan (Quevillon *et al.*, 2005).

*i*ProClass combines an integrated database with hypertext links to provide protein family, function and structure information (Wu *et al.*, 2004*a*). Sequences from PIR-PSD and UniProtKB are organised into superfamilies using PIR-SF. This organization forms the core of the database and functional annotations. Links to over 50 biological databases, literature and related sequences are included.

Certain integrated databases focus on the annotation of whole genomes. Such databases include the Protein Extraction, Description and Analysis Tool (PEDANT, Riley *et al.*, 2007), the Protein Centric Annotation System (PCAS Zhang *et al.*, 2003) and Predictions for Entire Proteomes (PEP, Carter *et al.*, 2003). PEDANT contains annotations for 468 proteomes, whereas PEP contains annotations for 105 proteomes. PEDANT and PEP include searches against SWISS-PROT, TrEMBL, the PDB and PROSITE. In addition, both PEDANT and PEP make the following predictions using existing methods: secondary structure, transmembrane regions, low complexity, signal peptides and coiled-coils. Searches against Pfam HMMs and SCOP PSSMs are further incorporated in the PEDANT annotation pipeline. Additional predictions in PEP include nuclear localization signals, functional class and long regions without regular structure (NORS).

Several genome annotation databases, dedicated to one specific organism, have been developed. Examples include PLASMODB (Stoeckert *et al.*, 2006) for *Plasmodium* species, the Mouse Genome Database (MGD; Blake *et al.*, 2006), and FlyBase (Gelbart *et al.*, 1997). PLASMODB offers various types of information for the genes from *P. falciparum*. Information for predicted proteins include: molecular weight (MW); isoelectric point (IP); the availability of a crystal or predicted 3D structure; similarity to proteins

in the PDB and protein sequence databases; InterPro domains; predictions of secondary structure, transmembrane helices, low complexity and signal peptides; protein-protein interactions and proteins exported to the host. Proteins predicted to be located in the mitochondrion are predicted by a neural network specific for *Plasmodium* species. Orthologous proteins in other *Plasmodium* species and paralogs are indicated for each protein. The results are displayed in a tabular format and an image of protein features is available.

In this chapter, an integrated structural annotation pipeline for *Plasmodium* species is presented. The focus of this annotation was to integrate a wide selection of structural feature annotations, including features not annotated in PLASMODB. Three species, *P. falciparum*, *P. vivax* and *P. yoelii* were annotated. However, this pipeline can easily be applied to other proteomes for structural annotation. These annotations will facilitate the identification of proteins suitable for structural studies such as X-ray crystallization, homology modeling and docking studies. Ultimately, the comparison of *P. falciparum* protein structures to human proteins will lead to effective therapeutic target selection. Each of the annotations included in the structural annotation system will be discussed in more detail.

For identification of proteins suitable for homology modeling, a template with known structure is necessary. Therefore, searches against the PDB are included. Although PSI-BLAST is more sensitive than BLASTP to identify distant homologs, it is computationally expensive. The time needed for a PSI-BLAST search increases exponentially with the length of the protein. Many *Plasmodium* proteins are very long. Therefore, a simple BLASTP search was performed against the PDB. Whereas the PDB contains whole sequences of known structure, SCOP contains domains within proteins of known structure. A HMMER search against SUPERFAMILY is therefore incorporated into the structural annotation system. To complement the BLAST-PDB and HMMER-SUPERFAMILY template searches, a group of proteins are subjected to threading. Threading makes use of a different approach than sequence-based searches and may therefore recognize

templates in cases where sequence-based searches are unsuccessful.

Following template recognition searches, the secondary structure is predicted for all of the proteins. Secondary structure may help identify the general class to which a protein belongs. In addition, secondary structure increases the accuracy of threading and disorder predictions. If a suitable template can be found for a protein, the secondary structure prediction can be used during homology modeling of the protein.

Transmembrane helices are predicted for the *Plasmodium* proteins in order to recognize unfavourable proteins for experimental structural determination. Transmembrane helices make proteins difficult to solubilize and crystallize. Other sequence features which make a protein unfavourable for crystallization, are long coiled-coil regions, regions of low complexity, long disordered regions and signal peptides. Therefore, these features are predicted for the *P. falciparum* proteins. Disordered regions are not visible in the X-ray of a crystal structure, since the regions are flexible and are not present in a single structural state. Long coiled-coils form regions of low complexity. Low complexity regions do not assume a globular shape and are therefore not readily crystallized. Signal peptides have similar characteristics to transmembrane regions and are therefore also difficult to crystallize. Signal peptides have an unfavourable influence on protein crystallization when the protein is very short and the signal peptide makes up a large proportion of the protein length. Other features making proteins unsuitable for protein crystallization includes post-translational modification sites, such as glycosylation sites. Therefore PROSITE motifs are identified in the *P. falciparum* sequences.

For the identification of proteins containing small molecule binding sites, SMID-BLAST searches are performed. SMID-BLAST (Hogue, 2006) is a tool for the annotation of small molecule interactions to proteins without a determined 3D structure. SMID-BLAST utilizes the RPS-BLAST program from the NCBI package to search for CDD domains, within a query protein sequence, known to interact with small molecules. Knowledge of

protein binding sites and the type of molecules that bind to the protein, are important for the identification of proteins suitable for docking studies.

To further prioritize proteins for possible structural studies, proteins of functional importance must be identified. For this reason, protein-protein interactions were annotated from literature (LaCount *et al.*, 2005). Proteins exported to the RBC may also indicate functional importance and were therefore annotated from literature. In addition, searches against Pfam and PRINTS identified functional domains.

Although many of the analyses performed are present in PLASMODB, certain features important for deciding on targets for structural studies are not available. PLASMODB does not include disorder and coiled-coil predictions. The annotation of small molecule binding sites is very important to identify proteins which bind to drug-like molecules and therefore to select suitable proteins for docking studies. The additional threading of proteins may also identify structural templates for proteins when sequence-based searches are not able to. Groups of proteins containing one predicted or known feature can be viewed in PLASMODB, however queries for selecting proteins containing two or more features are not available. In the structural annotation presented here, a tool to select groups of proteins containing one or more predicted features is available. This tool helps select proteins of interest for structural studies.

2.2. Methods

2.2.1. Pipeline construction

The pipeline was constructed in Python 2.4.3. A class was created for every analysis, all inheriting a central class responsible for submitting jobs to a Linux cluster. Every job is a qsub shell script containing the commands and options for the specific program being run with specified input and output files. An analysis for a specific sequence is performed on one of 64 2.4 Ghz CPUs on the deepthought cluster. Selected analysis can be done

by the administrative user. The administrative user must specify whether the analysis is being run for the first time on this specific species or if the results are being updated. The directory containing all the separate FASTA input files must also be specified. A standard FASTA parser is available to parse one FASTA file into separate FASTA files. Analysis can be extended to other proteomes and the database can optionally be updated by the administrative user. Figure 2.1 is a representation of the pipeline, database and web interface components.

2.2.2. Database

Many systems for creating and managing a database exist, including MySQL, Oracle and PostgreSQL. PostgreSQL is an open source SQL compliant relational database management system. It has the advantages of having programming interfaces for many programming languages including Perl, Python, C, C++ and JAVA. In addition, PostgreSQL runs on all major operating systems. The database for the structural annotation system was constructed in PostgreSQL, and contains tables for every analysis in the pipeline and for every species. The species tables contain sequence data and descriptions, with indexes on the sequence names and species columns (Figure 2.2).

2.2.3. Web interface

The web interface was constructed using ZOPE as a web application managing platform. Application servers such as ZOPE allow the managing of a web site, presentation of dynamic content and the integration of diverse systems such as files, relational databases and separate web sites. ZOPE runs on most popular operating system platforms: Linux, WindowsNT/2000/XP, Solaris, FreeBSD, NetBSD, OpenBSD, and Mac OS X. ZOPE can also be extended using the Python scripting language. Using ZOPE and associated Python external methods, the following features were incorporated into the web interface: a keyword search; a PLASMODB ID-search; a protein selection tool; links to

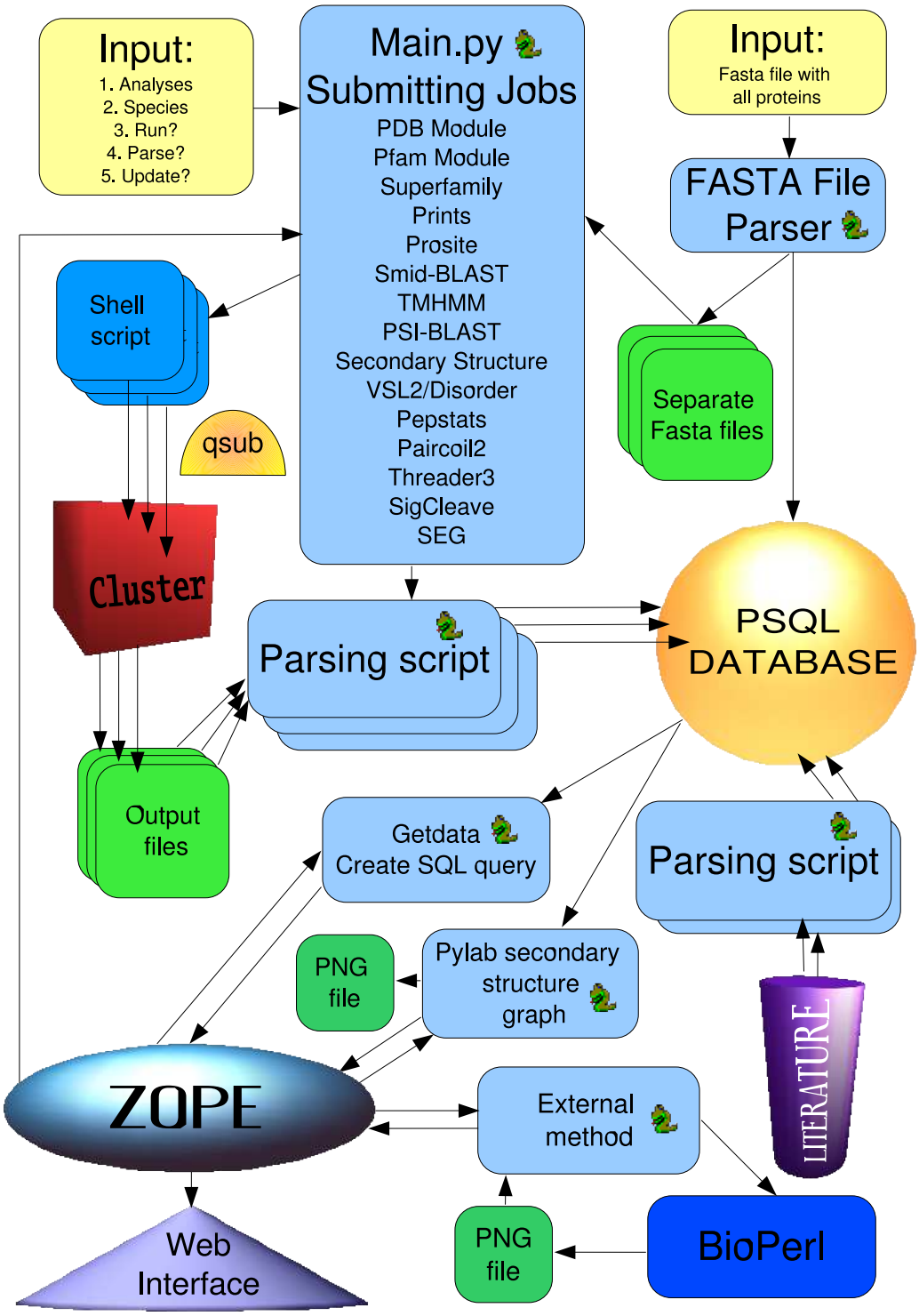


Figure 2.1: A flow diagram of the components of the structural annotation system and interactions between them.

Figure 2.2: A diagram of information stored in the database.

PLASMODB, Pfam, PDB, Superfamily, PRINTS, PROSITE, Superligands, interacting proteins and *Plasmodium* orthologs; a sequence feature image; and secondary structure and protein disorder confidence graphs.

2.2.4. Tools in the annotation system

In the following section, the focus will be on the specific analyses incorporated into the pipeline. Most of the analyses are independent of the pipeline and can optionally be invoked. Certain analyses such as secondary structure prediction, threading and disorder predictions are dependent on PSI-BLAST searches and must be run in the correct order with respect to each other.

Protein statistics

The EMBOSS 3.0.0 program Pepstats was used to calculate general protein statistical features from the sequence such as composition frequencies, the molecular weight, iso-electric point (IP), extinction coefficient (EC) and average residue weight (ARW).

Structure template recognition

Three approaches were included for the purpose of template recognition:

1. BLAST against the PDB

   The Blastall program from the NCBI BLAST package, version 2.2.10 was run against the PDB. A cutoff e-value of 20 was used, in order to rank proteins according to similarity in the PDB.
2. HMMER against Superfamily

   Hmmpfam from the HMMER package version 2.3.2 was run against the Superfamily database. A cut-off e-value of 1e-01 was used.
3. Threader3

   Threading was done on selected sequences by THREADER3. As input, a secondary structure prediction output file from Psipred was used. Only sequences shorter than 400 amino acids were used. The first reason is that known structural domains are

rarely longer than 400 amino acids, therefore predictions for longer sequences would be unreliable. The second reason is that longer sequences are computationally very expensive to run, and a sequence between 500 and 800 amino acids can take between six to twelve hours to run on a reasonably fast CPU. The Threading Expert program is part of the Threader3 package and attempts to calculate a single value for judging the accuracy of the prediction. In the Threader3 output file, there are 14 columns, many of which contribute in different ways to the decision of which model is correct. When the Threading Expert is used to process the Threader results, the secondary structure prediction file has to be used as input for Threader. The Threading Expert was used to interpret the results of Threader3 automatically for all the sequences predicted.

Transmembrane domains

Transmembrane helix predictions were made by TMHMM2. TMHMM is an accurate transmembrane region predictor although the topology might be wrong. Therefore, TMHMM should not be used for predicting the orientation with regard to the inside and outside of the membrane (Krogh *et al.*, 2001). TMHMM2 is also easily implemented and simple to use. The program requires an input file in FASTA format containing the query sequence and produces an output file containing the topology of the transmembrane regions in the following format: i172-191o201-220i233-255o294-316i. Thus, amino acids 1-171 are on the inside of the membrane, amino acids 172-191 form a transmembrane helix, amino acids 192-200 form an outside loop, etc.

Signal peptides

The EMBOSS 3.0.0 program, Sigcleave, was used to predict signal peptides. SigCleave is not the most accurate program for prediction of signal peptides, being one of the first methods developed but is freely available and easily implemented.

Coiled-coils prediction

Coiled-coils were predicted by Paircoil2. Paircoil2 uses as input a FASTA file of the query sequence and generates an output file containing a table of prediction scores and the position in the heptad repeat pattern (abcdefg) for each amino acid. The more heptad repeats in a coiled-coil, the more accurate the prediction. A p-score of 0.03 corresponds to a sensitivity of 0.73 and a specificity of 0.998 (McDonnell *et al.*, 2006). A coiled-coil structure was assumed if the p-score was less than 0.025.

Secondary structure prediction

Secondary structure predictions were made by Psipred version 2.5. PSI-BLAST, from the NCBI BLAST package blast-2.2.12-ia32-linux was run on the sequences. Three iterations as recommended by Psipred, were used and a cutoff e-value of 0.001. The -Q option is used to generate an ASCII formatted PSSM. In addition, a byte-encoded checkpoint file is generated (using the -C option) to store the query and frequency count ratio matrix, which can later be reused. Makemat, a program in the IMPALA package, was used to convert this non-human readable file to a binary formatted file. Psipred makes use of the binary file to reduce loss of precision when parsing the original ASCII position specific matrices. Psipred also takes a FASTA file of the query sequence as input, and makes two final output files. One file is in a horizontal format showing the amino acid and its position, the prediction, and the confidence level. The other output file contains a complete table of results which shows the individual coil, helix, and strand network outputs. The output files contain a one-line header to allow THREADER to automatically recognize these files.

Disorder predictions

Protein disorder was predicted by the Disprot/VSL2 predictors. The program invokes different predictors depending on the options used. Options used correspond to input files given to the program. The most accurate predictions result from using a PSI-BLAST

PSSM together with Psipred- and PHDsec secondary structure prediction files (Table 2.1). Options for the PSSM and the Psipred files are used. As with the Paircoil2 output file, a table is generated containing per residue confidence scores for the prediction.

Table 2.1: The influence of different input files on the prediction accuracy of VSL2. The highest accuracy results from using a PSSM, and secondary structure predictions from PHD and Psipred. Extracted from Peng *et al.*, 2006. AA refers to the amino acid sequence.

AA	**PSSM**	**PHD**	**PSI**	**Accuracy**
+				78.6
+		+		79.1
+			+	79.9
+		+	+	80.1
+	+			80.7
+	+	+		81.2
+	+		+	81.4
+	+	+	+	81.6

SMID-BLAST

Version 1.02 of SMID-BLAST was run with all the default settings. No e-value or molecules per binding site limits were set. This allows for a maximum number of matches and users can decide which matches are significant based on the ligand scores. The input file is a FASTA file with any amount of sequences.

Functional family assignment

Hmmpfam from the HMMER 2.3.2 package was run against the Pfam_fs version 20 HMM file. The recommended cut-off e-value was set at 0.02 and output alignments were limited to the best five hits.

Motifs

1. PRINTS search

   A search against the PRINTS database was done by running the EMBOSS 3.0.0 program Pscan.

2. PROSITE search

   The EMBOSS 3.0.0 program Patmatmotifs was run on every sequence to search for motifs contained in the PROSITE database.

Protein-protein interactions

P. falciparum protein-protein interactions determined by high-throughput yeast-two hybrid experiments by La Count *et. al.* (2005), were annotated to the sequences. The supplementary tab-delimited data files were parsed to the database.

Exported proteins

Proteins predicted to be exported to the red blood cell, determined by the presence of the SS-signal and Pexel/VTS motifs, were annotated from supplementary material from Miller *et. al.* (2004) and Marti *et.al.* (2004).

2.2.5. Parsing the output

For every analysis a parsing script was written to insert output data into the database. This parsing script can be run automatically if indicated by the administrative user. These parsing scripts are specific for sequences with FASTA title lines in the format of PLASMODB release 5 protein sequence files. Each parsing script attempts to make interpretation of the specific results easier for the user. For example, parsing a Paircoil2 output file will calculate the amount of heptad repeats for a particular coiled coil region. Parsing TMHMM2 output, will calculate the amount of transmembrane segments, which in turn gives an idea of the specific class of proteins.

2.3. Results

2.3.1. Addition of an analysis

Analyses are easily added to the structural annotation system. An extra class must be added to the main script, SAMP.py. A table in the database must be created for the

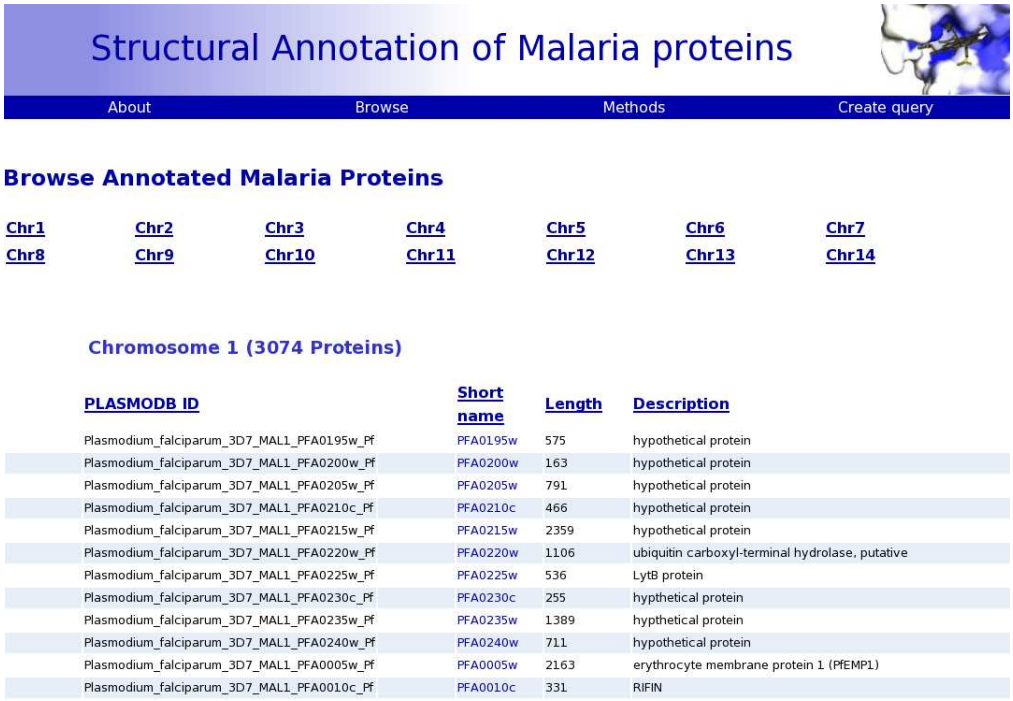


Figure 2.3: Browsing view of the web interface.

analysis and a parsing script must be written in order to add the results to the database. Web display and visualization of the results are not automatic. Extending the structural analysis to other species is easily implemented and requires only a directory name with separate FASTA files of each sequence. The database can be updated by entering '1' for both the run and update options of SAMP.py.

2.3.2. Web Interface

Sequence search

Sequences can be searched for by keywords, by browsing proteins on a chromosome and by designing complicated queries according to user specifications. Figure 2.3 shows how predicted proteins for each chromosome can be browsed and Figure 2.4 illustrates how proteins can be searched for by keyword or PLASMODB ID.

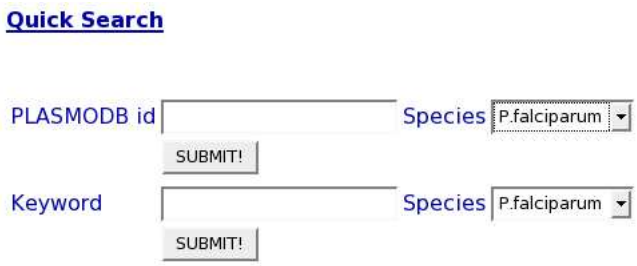


Figure 2.4: Sequences can be searched for by either their PLASMODB ID or a keyword.

Create a query

PostgreSQL queries are constructed piece by piece from check boxes ticked by the user in a Python script added to ZOPE as an external method (Figure 2.1). In the database, the summary table (proteome_analysis) contains pivotal annotations for the selection of groups of proteins having specified annotations. For example, the table contains e-values, sequence identity and sequence coverage of PDB matches and the percentages of coiled-coils, disorder, signal peptide and transmembrane regions within the query sequence. Furthermore, the table contains annotations of family domains and motifs, the amount of interactions and transmembrane helices and proteins which have a crystal structure of homology model. For the query tool, a PostgreSQL query is constructed as follows: `'SELECT COLUMNS FROM proteome_analysis WHERE'` is concatenated to criteria set by each checked box with `'AND'` in between. Figure 2.5 shows how a query can be made by a user through the web interface. Querying from only one table, ensures that many criteria can be set and that the execution time is fast. This allows for the selection of all proteins with no transmembrane, disordered, coiled-coils regions and setting a cut-off e-value for PDB matches. Such a selection is useful for identifying suitable targets for experimental structure determination. In addition, the amount of transmembrane helices and interactions can be set as well as the presence of a Pfam domain, PRINTS motif, export motif and PROSITE motif.

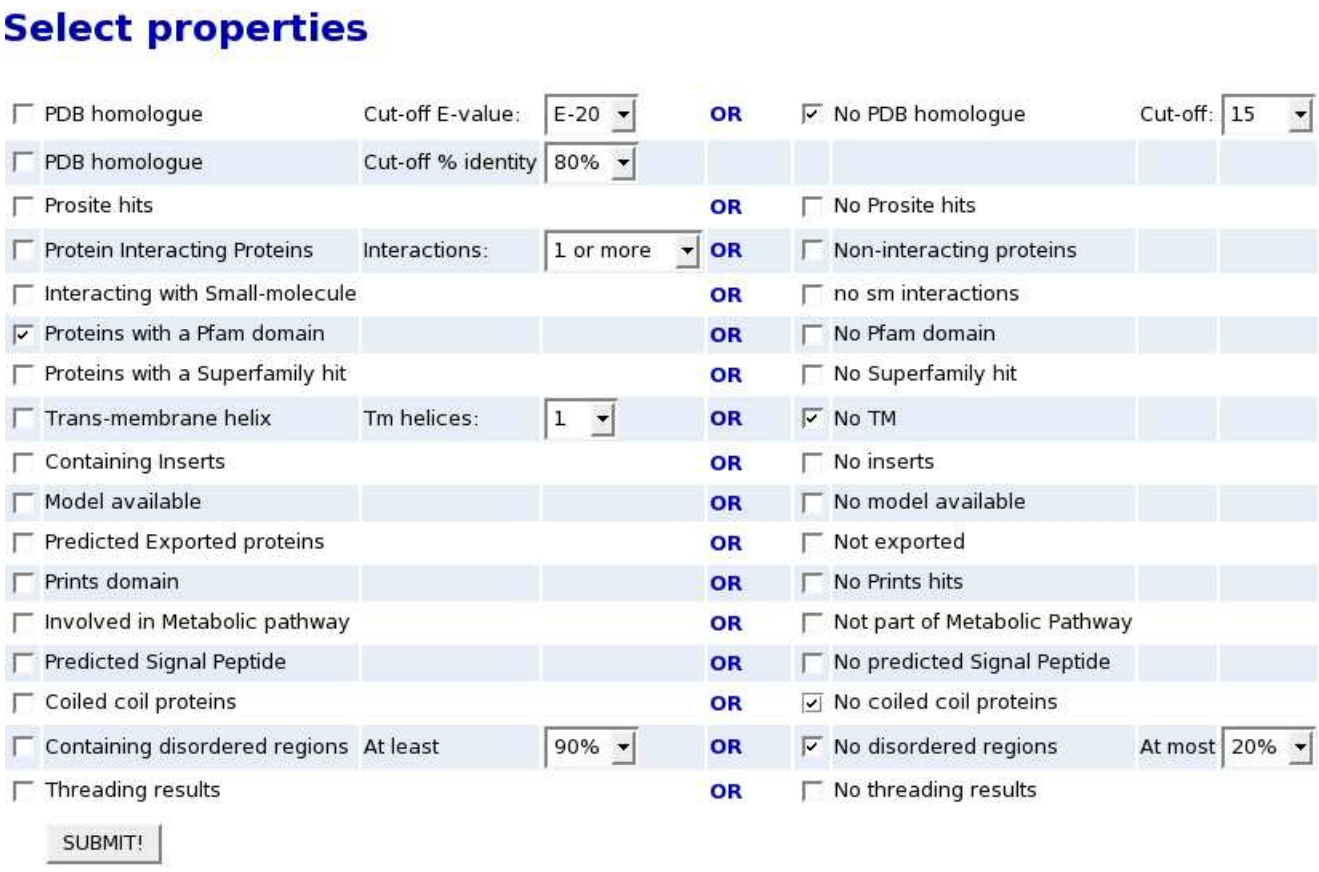


Figure 2.5: Groups of proteins can be selected by the 'Create query' tool.

Results page

Figures 2.6 to 2.14 display different parts of the results page for different sequences. The results page starts with some typical sequence statistics as calculated by Pepstats. The next section provides the user with a summary image displaying database coverage, motifs, disordered regions, coiled-coils, low complexity and transmembrane helices (Figure 2.8). Each of the following sections lists the results of a specific analysis. The results include links to databases, start and end positions on the query sequence, scores, e-values and descriptions.

Following the sequence length and PLASMODB sequence description, is a colour-coded predicted protein sequence. Colours indicate the amino acid type (Figure 2.6). Protein statistics calculated by Pepstats include the molecular weight, average residue weight, charge, IP, EC, molar extinction coefficient (MEC) and improbability of expression in inclusion bodies (IEIB). Percentages of each of the amino acids in the query sequence are given in addition to percentages of types of amino acids (tiny, small, aliphatic, aromatic, non-polar, charged, basic and acidic).

A summary of the predicted sequence features for putative uridine phosphorylase (PFE0660c) is shown in Figure 2.7. In the summary image, the best PDB and Superfamily hits based on the score are illustrated. All the Pfam hits above a cut-off are displayed since the protein might have more than one non-overlapping functional domain. Additional PDB and Superfamily hits can be viewed by following the links. Links to the PDB, Superfamily and Pfam are available. Figure 2.8 shows the output displayed for the BLAST-PDB search. Similar results are displayed for Superfamily and Pfam. The PDB descriptions are sometimes partial as a result of the BLAST output file which only allows for a maximum amount of characters.

A graph constructed by Matplotlib displays the confidence values for helix, strand and disorder over the length of the protein sequence (Figure 2.9). The graph also shows

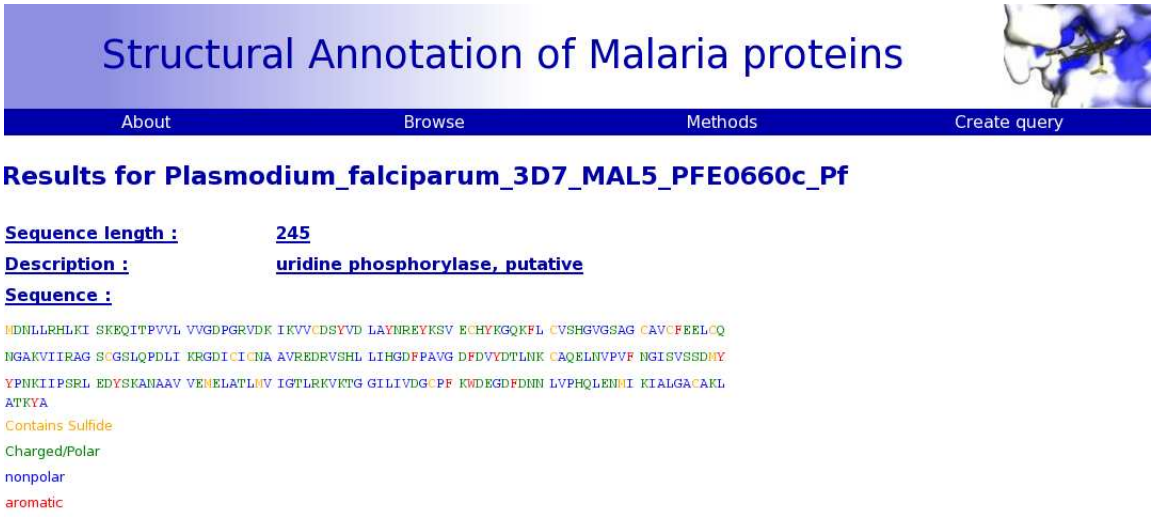


Figure 2.6: General sequence statistics for PFE0660c.

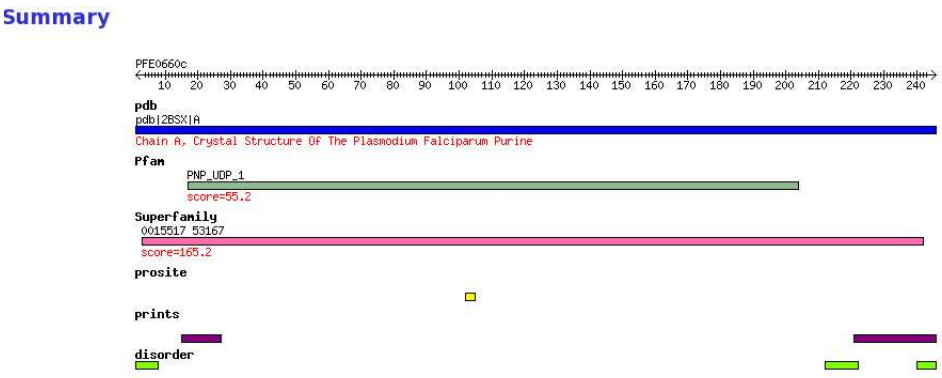


Figure 2.7: Graphic representation of predicted sequence features for PFE0660c, a putative uridine phosphorylase.

Blast PDB hits

PDB id:	E-value:	Score:	% Coverage:	Identities	Description
pdb|2BSX|A	1e-143	502.0	99.5918367347	(245, 245)	Chain A, Crystal Structure Of The Plasmodium Falciparum Purine
pdb|1Q1G|F	1e-143	501.0	99.5918367347	(244, 245)	Chain F, Crystal Structure Of Plasmodium Falciparum Pnp With 5'-
pdb|1SQ6|A	1e-140	491.0	99.1836734694	(240, 244)	Chain A, Plasmodium Falciparum Homolog Of Uridine
pdb|2B94|A	1e-119	422.0	98.7755102041	(200, 243)	Chain A, Structural Analysis Of P Knowlesi Homolog Of P Falciparum
pdb|1T0U|B	5e-22	99.8	95.1020408163	(70, 249)	Chain B, Crystal Structure Of E.Coli Uridine Phosphorylase At 2.2 A
pdb|1U1G|F	5e-22	99.8	95.1020408163	(70, 249)	Chain F, Structure Of E. Coli Uridine Phosphorylase Complexed To 5-
pdb|1ZL2|F	7e-22	99.4	95.1020408163	(71, 249)	Chain F, Crystal Structure Of The Uridine Phosphorylase From
pdb|1LX7|B	7e-21	95.9	68.1632653061	(56, 170)	Chain B, Structure Of E. Coli Uridine Phosphorylase At 2.0a
pdb|1Z39|A	1e-17	85.5	90.2040816327	(72, 228)	Chain A, Crystal Structure Of Trichomonas Vaginalis Purine
pdb|1ODI|F	2e-15	77.8	80.8163265306	(57, 201)	Chain F, Purine Nucleoside Phosphorylase From Thermus Thermophilus

View image of 5 best hits

Figure 2.8: Display of BLAST-PDB results.

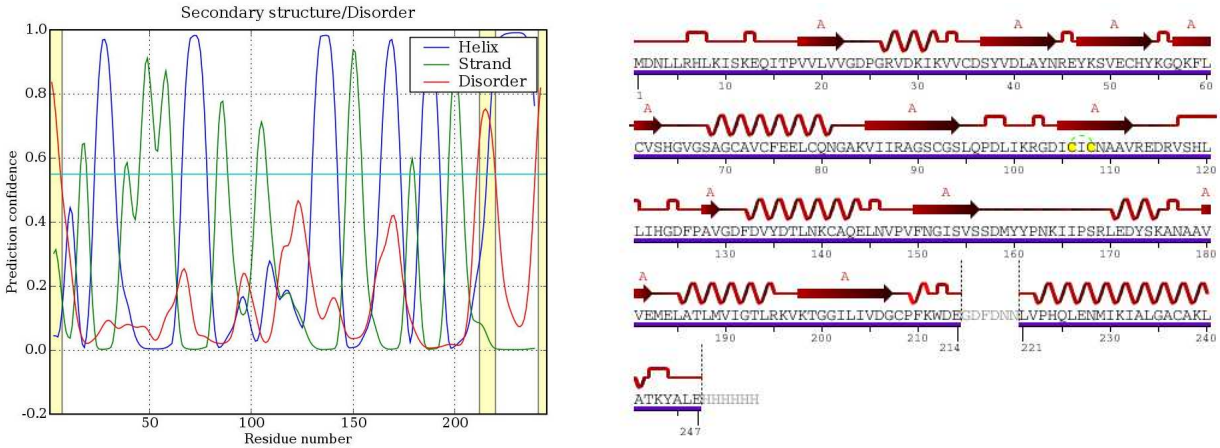


Figure 2.9: (a) Graph depicting confidence values for predicted helix, strand and disorder. (b) DSSP secondary structure assignment for 2BSX, the crystal structure of *Plasmodium falciparum* Purine nucleoside phosphorylase (`http://www.rcsb.org/pdb/explore/sequence.do?structureId=2BSX`).

a cut-off value of 0.55, below which disorder should not be assumed. The yellow coloured areas are regions of predicted disorder above the cut-off. A graphical representation of the DSSP assignment (Dictionary of protein secondary structure; Kabsch and Sander, 1983) for the PDB structure 2BSX was obtained from the PDB and is also displayed in Figure 2.9. The prediction and assignment correlate across the length of the sequence. The Psipred secondary structure and VSL2 disorder prediction output files are included in Appendix A. Secondary structure was 85.3% correctly predicted, not including the disordered stretch (215-220). A three-amino acid helix (209-211) and a two-amino acid strand (128-129) were both mispredicted as coils with confidence values of 0.6 and higher. High confidence (>0.6) disordered regions were predicted for amino acids 1-6, 213-220 and 241-245. The total predicted disorder is 7.9%. The crystal structure contains 2.4% disorder.

Regarding Threader results, the most important columns for manual interpretation are displayed. Figure 2.10 displays the Threader results for PFE0660c. Links to the particular PDB structures are also included. The threading expert ranks the ten best alignments. The Z-weighted energy refers to the Z-score of the weighted sum of pairwise and solvation energies. This is the most important column for deciding on a template. According to the Threader3 manual, Z-scores above 4.0 are significant, Z-scores between 3.5 and 4.0 are less reliable and Z-scores between 2.7 and 3.5 are unreliable. Z-scores below 2.7 are regarded as poor. The template energy refers to the native energy of the known structure, and the Threading Energy column refers to the raw pairwise energy sum for the particular threading. The developers report that in their experience, if the threading energy is lower than that of the native structure, the match is probably false. The number of aligned residues-column is another column to consider; the higher the amount of residues, the better the alignment. It is of value to examine at the percentage structure aligned, since partial matches to structural folds are often incorrect.

Figure 2.11 reports the protein-protein interactions for cholinephosphate cytidylyl-

THREADER

PDB id:	Threading Expert	Z-weighted Energy	Template Energy	Threading Energy	No aligned residues	Percentage Structure	Percentage Sequence	Score
1ecpA0	0.932024	5.71	-800.09	-987.37	232.0	97.9	94.7	370.8
1nw4A0	0.923265	6.11	-992.39	-1136.61	230.0	95.1	94.3	404.4
1jdsA0	0.919071	5.49	-762.17	-941.84	223.0	98.7	91.0	341.6
1b8oA0	0.901426	4.14	-830.9	-785.01	233.0	83.2	95.1	301.7
1jysA0	0.886631	4.59	-897.68	-914.43	216.0	94.3	88.6	352.8
1k3fA0	0.875137	3.96	-643.85	-839.44	221.0	87.7	90.6	353.9
2dri00	0.844536	3.19	-679.86	-613.01	210.0	77.5	85.7	18.1
1dhpA0	0.835161	2.7	-769.98	-433.94	205.0	70.2	83.7	14.5
1qr7A0	0.833801	3.64	-1044.15	-763.95	220.0	65.4	90.2	-36.3
2tysA0	0.820411	2.29	-974.92	-492.84	206.0	80.8	84.1	-15.7

Figure 2.10: Display of Threader results.

Interactions

Bait	Bait Start	Bait End	Prey	Description	Prey Start	Prey end
MAL13P1.86	1567	2060	PF10_0150	"methionine aminopeptidase, putative "	345	646
MAL8P1.19	2726	2954	MAL13P1.86	"hypothetical protein, conserved "	2412	2685
PF13_0213	140	424	MAL13P1.86	"60S ribosomal subunit protein L6e, putative "	1473	1749
PFE1590w	331	521	MAL13P1.86	early transcribed membrane protein	2412	2598
PFI0875w	328	596	MAL13P1.86	heat shock protein	115	408
PFL0130c	3759	4197	MAL13P1.86	"hypothetical protein, conserved "	1552	1749
PFL2345c	113	332	MAL13P1.86	tat-binding protein homolog	1851	2151

Figure 2.11: Table of protein-protein interactions as annotated from literature.

transferase (MAL13P1.86). Descriptions and links to the output reports of the interacting proteins are displayed.

For transmembrane regions, disorder and interactions with small molecules, putative cAMP-specific 3,5-cyclic phosphodiesterase (MAL13P1.118) will be used as an example. The summary image for MAL13P1.118 is shown in Figure 2.12. Generally, the binding positions of 5 to 10 of the best-scoring molecules are graphically represented by red triangles within the summary image. Figure 2.13 illustrates the listing of predicted interactions with small molecules in the results page together with links to PDB, CDD, PubChem and Superligands (Michalsky *et al.*, 2005). Superligands provides information on the small molecule's similarity to known drugs.

Transmembrane helices are shown as orange blocks in the summary image. The

Summary

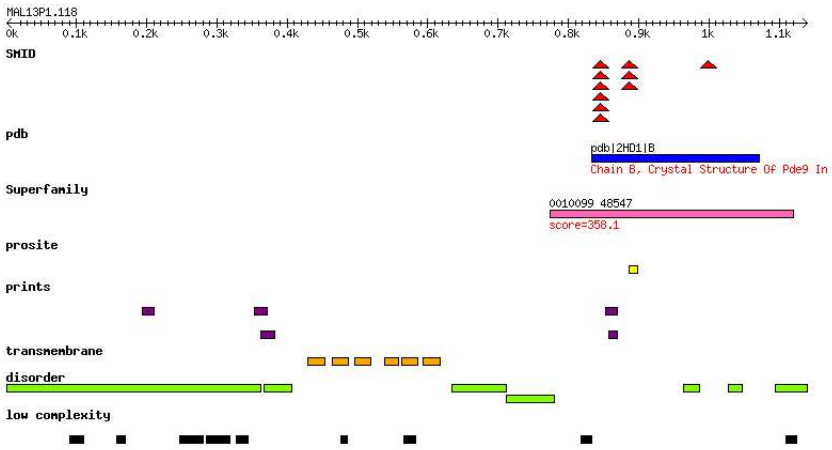


Figure 2.12: Summary image for putative cAMP-specific 3,5-cyclic Phosphodiesterase 4D.

Interactions with small molecules

Molecule	Ligand Score	B-site no.	Locations
7DE (7DE)	134.218	1	845-846,999-1000,1013-1014,957,959,1002,1010,1017,1021
8BR (8BR)	160.971	1	845-846,999-1000,850,959,1002,1010,1017,1021
AMP ([[(2R,3S,4R,5R)-5-(6-amino-9H-purin-9-yl)-3,4-dihydroxytetrahydrofuran-2-yl]methyl dihydrogen phosphate)	331.677	1	845-846,886-887,999-1000,850,959,1002,1017,1021
GMP ([[(2R,3S,4R,5R)-5-(2-amino-6-oxo-1,6-dihydro-9H-purin-9-yl)-3,4-dihydroxytetrahydrofuran-2-yl]methyl dihydrogen phosphate)	290.336	1	999-1000,846,850,887,890,1021
Mg2 (Magnesium)	557.113	1	886-887,916,919,957,999
Mn2 (manganese(2+))	140.86	1	887,957
PIL (PIL)	116.867	1	845-846,999-1000,1013-1014,1017-1018,957,959,1002,1021
ROL (ROL)	107.208	1	845-846,999-1000,1013-1014,1017-1018,959,1002,1008,1021
ZAR (ZAR)	198.545	1	845-846,999-1000,850,887,1002,1017,1021
Zn2 (Zinc)	9033.119	1	886-887,850,999

References

Molecule	MOD	BIND	PDB	CDD	PubChem.SID
7DE	23004	308749,308748,308750	1Y2J 1Y2K	28958	
8BR	15403	305893,305892	1RO9	28958	
AMP	1043	251562,251560,251561,256757,256758	1PTW 1TB5 1TB7 1ROR	28958	3322
GMP	1602	256753,256752	1T9S	28958	3444
Mg2	166	232345,245460,254036,255540,256256	1F0J 1MKD 1RKP 1SO2 1SOJ	28958	
Mn2	195	305889	1RO6	28958	841106
PIL	24567	306647,306847	1XM4 1XON	28958	
ROL	24732	248332,248331,248330,256760,305891	1Q9M 1OYN 1TBB 1RO6	28958	
ZAR	14849	245461,306850	1MKD 1XOR	28958	
Zn2	172	248329,232344,245459,254035,256255	1Q9M 1OYN 1PTW 1RO9 1ROR	28958	841092

Figure 2.13: Small molecule interactions as predicted by SMID (MAL13P1.118).

Similarity with other Plasmodium species

Species	Protein	E-value:	Score:	% Coverage	Identities
chabaudi	PC_RP0736PC000630.00.0	1e-101	362.0	99.1836734694	(172, 244)
yoelii	MALPY01425PY04622	1e-113	403.0	99.1836734694	(192, 244)
berghei	PB_RP3406PB000500.03.0	1e-115	407.0	99.1836734694	(193, 244)
vivax	ctg_6840Pv080575	1e-118	418.0	97.1428571429	(199, 239)

Figure 2.14: Links to similar proteins from *P. chabaudi, P. yoelii, P. berghei* and *P. vivax*

cAMP-specific 3,5-cyclic phosphodiesterase is predicted to have 6 transmembrane helices. Predicted disordered regions are shown as green rectangles in the summary image. Images showing transmembrane helices and disordered regions respectively, are also available in the results page. Orthologs from *P. yoelii, P. berghei, P. vivax* and *P. chabaudi* are displayed for species comparison (Figure 2.14). In addition, proteins which are exported to the RBC are reported. At the bottom of the page is the corresponding PLASMODB link.

2.3.3. Validation study for PFE0660c

Putative uridine phosphorylase (PFE0660c) will be discussed with regard to the validation of BLAST-PDB, Threader, Secondary structure prediction, disorder prediction, HMMER-Superfamily and *Plasmodium* ortholog results. Refer to Figures 2.7 to 2.10.

BLAST-PDB

The best PDB hit for PFE0660c has the PDB code 2BSX chain A. This PDB code corresponds to the crystal structure of *P. falciparum* purine nucleoside phosphorylase (PNP) in complex with inosine (Schnick *et al.*, 2005). The sequence similarity is 100%. The next three PDB hits correspond to *P. falciparum* PNP in complex with different compounds or *Plasmodium* PNP homologs. A *P. falciparum* PNP crystal structure in complex with sulfate and immucillin-H (PDB code 1NW4) was not included in the results. The remaining hits are the crystal structures of PNP from more distantly related organisms with approximately 30% sequence similarities. BLASTP thus correctly identified

4 of the 5 most similar sequences in the PDB to PFE0660c, and furthermore correctly identified more distantly related PNP.

Threader

Threader correctly identified PNP from *E.coli* (PDB code 1ECP) and *P. falciparum* (PDB code 1NW4) as templates for PFE0660c. The *E.coli* PNP scored a higher ranking by Texp. However, 1NW4 had a better Z-score. Interestingly, 1NW4 is the protein structure that was not included in the BLAST-PDB results. None of the other *P. falciparum* PNP structures were identified by Threader. The remaining templates, identified by Threader, include Bovine PNP (PDB code 1B8O), 5'-methylthioadenosine phosphorylase (MTA; PDB code 1JDS), MTA/S-adenosylhomocysteine nucleosidase (PDB code 1JYS) and *E.coli* Uridine Phosphorylase (PDB code 1K3F). These templates are part of the Purine and Uridine Phosphorylase superfamily and had Z-scores between 5.49 and 3.96. Templates with lower Z-scores (2.29-3.64) belonged to the aldolase superfamily and had a TIM beta/alpha-barrel fold. The templates with Z-scores between 2.29 and 3.64 falls in the 'unreliable' range of Z-scores, as defined by the developers of Threader and the Z-scores for templates containing the phosphorylase/hydrolase fold correspond to the 'reliable' range of Z-scores.

Secondary structure and disorder

A comparison of the predicted secondary structure for PFE0660c and the assigned secondary structure for 2BSX showed that 85% of the amino acids were correctly predicted as either helix, strand or coil. A short 3-10 helix and a short beta strand were incorrectly predicted as coils with confidence values between 0.6 and 0.9. In addition, a short alpha helix and a short beta strand were incorrectly predicted in the place of a turn and a coil, respectively. The mispredicted helix and strand had confidence scores in the range of 0.3 to 0.6 and 0.1 to 0.2, respectively. Finally, the H-state was predicted for two amino acids situated on the border of a correctly predicted beta strand with confidence values of 0

and 0.1. Conclusively, predictions of short (two to three amino acids) helices and strands are not reliable. Transitions from H to E and E to H predicted states are not physically plausible and are incorrect. Furthermore, prediction confidence values between 0.0 and 0.4 are highly unreliable.

Assuming that only the greyed-out area in the secondary structure assignment reflects true disorder, the structure contained 2.4% disorder and the first and last predicted disordered regions were not correct. However, all the truly disordered amino acids were correctly identified.

Superfamily and family assignment

All the reported Superfamily hits correspond to different HMMs of the Purine and Uridine Phosphorylase superfamily containing the Phosphorylase/hydrolase-like SCOP fold. The protein family for PFE0660 were correctly identified by HMMER as the Phosphorylase family which includes PNP, Uridine phosphorylase and 5'-methylthioadenosine phosphorylase (MTA). No motifs associated with the phosphorylase family were included in the PROSITE or PRINTS results. A scan of PROSITE identified a 3-amino acid motif, RGD, associated with cell adhesion. This motif has a high probability of occurrence. A PRINTS search identified one element from two fingerprints, respectively. The first element corresponded to a class 4 match, with only one out of six elements present, associated with the Insect alcohol dehydrogenase signature. The second motif was one of six elements from the Ribonucleotide reductase large chain signature. Hence, both the PROSITE and PRINTS searches were unsuccessful in identifying the family of PFE0660c.

Plasmodium orthologs

Four *Plasmodium* orthologs of PFE0660c were identified using BLASTP searches. Three sequences from *P. chabaudi* (PC000630.00.), *P. berghei* (PB000500.03.0) and *P. vivax* (Pv080575) respectively, are all described as putative uridine phosphorylase. PC00030,

PB000500.03.0, and Pv080575 had 70%, 79% and 83% identity to PFE0660c respectively. Furthermore, a sequence from *P. yoelii* (PY04622) with 79% identity are described as a putative purine nucleoside phosphorylase. Thus, four *Plasmodium* orthologs were successfully identified.

2.3.4. Validation study for MAL13P1.118

Putative cAMP-specific 3,5-cyclic phosphodiesterase (MAL13P1.118) will be discussed as an example of predictions of small molecule interactions, transmembrane helices and domain boundaries. Refer to Figures 2.12 to 2.13.

Small molecule interactions

Putative cAMP-specific 3,5-cyclic Phosphodiesterase 4D (cAMP-PDE4D; MAL13P1.118) is predicted to bind to AMP as expected. The protein is also predicted to bind to 8-bromo-AMP, GMP, Mg^{2+}, Zn^{2+}, Mn^{2+}, Zardaverine (ZAR), Rolipram (ROL) and Piclamilast (PIL). The proteins in complex with the small molecules and their PDB codes, corresponding to the four best scores are listed in Table 2.2. Most of the proteins bound to these molecules correspond to human cAMP-PDE4D. Other compounds found in complex with human cAMP-PDE4D, which were not predicted to bind to MAL13P1.118, include Cilomilast (1XOM), Roflumilast (1XOQ), 1,2-Ethanediol (1Y2B) and 3-Isobutyl-1-Methylxanthine (1RKO).

Transmembrane helix predictions

Six transmembrane helices were predicted for MAL13P1.118. A BLASTP search of MAL13P1.118 identifies two other phosphodiesterases overlapping in the predicted transmembrane area (not shown). These proteins (DdPDE4, a cAMP-specific phosphodiesterase from *Dictyostelium discoideum* and TcPDE1, a cAMP-specific phosphodiesterase from *Trypanosoma cruzi*) are membrane-bound as found by Bader *et al.*, 2006

Table 2.2: Experimentally determined proteins in complex with molecules predicted to bind to MAL13P1.118

PDB code	Protein	Species	Molecules bound
1RO9	cAMP-PDE4B	*H.sapiens*	Zn^{2+}
1PTW	cAMP-PDE4D	*H.sapiens*	AMP, Zn^{2+}
1TB5	cAMP-PDE4B	*H.sapiens*	AMP
1TB7	cAMP-PDE4D	*H.sapiens*	AMP
1ROR	cAMP-PDE4B	*H.sapiens*	AMP, Zn^{2+}
1FOJ	cAMP-PDE4B	*H.sapiens*	Mg^{2+}
1MKD	cAMP-PDE4D	*H.sapiens*	Mg^{2+}, ZAR
1RKP, 1SO2, 1SOJ	cGMP-PDE3B	*H.sapiens*	Mg^{2+}
19QM, 1YON	cAMP-PDE4D	*H.sapiens*	Zn^{2+}
1XOR	cAMP-PDE4D	*H.sapiens*	ZAR

and D'Angelo *et al.*, 2004. In addition, it was found that mouse and human PDE3 are localized on the endoplasmic reticulem (ER) membrane, containing six transmembrane helices (Shakur *et al.*, 2000). Both the C-terminal and N-terminal domains are on the cytoplasmic side (outside) of the ER membrane. TMHMM2 predicted the topology for MAL13P1.118 as o430-452i464-486o496-518i539-558o563-585i594-616o, thus both the N-terminal and C-terminal domains are predicted to be on the outside of the membrane corresponding to the previous finding for human and mouse PFE3.

Domain boundary determination

The transmembrane and disorder predictions can be used together with the PDB and Superfamily domains to identify domains in the protein. The first domain is extracellular and streches from amino acids 1-429. This domain is predicted to contain a high percentage of disorder (84% predicted with a confidence of 0.93). The six-transmembrane helix domain follows, stretching from amino acids 430 to 616. The last domain is globular and is predicted to bind to AMP, Zn^{2+} and Mg^{2+}. Psipred predicts 14 alpha-helices with confidence values higher than 0.6. These helices correspond to the helix content of the PDB entry for human phoshodiesterase 9 (PDB code 2HD1). This protein contains 14 long (6 amino acids or more) alpha-helices. Additionally, the crystal structure contains three alpha-helices shorter than 5 amino acids and five 3-10 helices of three amino acids

each.

2.4. Discussion

Secondary structure predictions by Psipred correlate in general with the assigned secondary structures from the crystal structure of proteins. Secondary structure predictions are more accurate for longer alpha-helices and beta-sheets than for short RSS. Turns are often predicted as small disordered regions by VSL2. Although VSL2 takes Psipred secondary structure predictions as input, high confidence disorder predictions sometimes overlap with high confidence RSS predictions by Psipred. Predicted short disordered regions have lower confidence scores than predicted long disordered regions, and overlap with RSS predictions more often than predicted long disordered regions. Disorder predictions seem to have a higher sensitivity than specificity, although it must be kept in mind that intrinsically disordered regions may become structured upon phosphorylation and the binding of molecules. Therefore, predictions of disorder may seem to overpredict disorder when it is compared to structures of similar sequences in the PDB which are complexed with stabilizing molecules. Also, different types of disorder exist, and not all types of disorder are shown as 'greyed-out' areas in the crystal structure (Vucetic, Brown, Dunker and Obradovic, 2003).

Threader correctly predicts templates for sequences when Z-scores above 4.0 are obtained. In some cases, Threader predicts suitable templates which are not detected by BLAST-PDB. In the majority of cases, BLAST-PDB and HMMER-Superfamily searches are more effective for fold recognition. The HMMER search against Superfamily detects similarities in a queried sequence to a SCOP fold when no PDB hits can be found with BLAST. Therefore, the HMMER-Superfamily search is more sensitive than a BLAST-PDB search. However, a HMMER-Superfamily search takes much longer to complete than a BLAST-PDB search.

PRINTS, PROSITE and Pfam searches all contribute to the identification of functional families. Examples for family identification by PRINTS include PF14_0598 and PFL0705c. For these two proteins, Pfam and PROSITE searches did not detect similarities to families within these databases. A HMMER-Pfam search identified the family of PF10_0209. Searches against PROSITE and PRINTS did not give meaningful results. For MAL13P1.118, the correct family was identified from Pfam and PROSITE. A PRINTS scan did not recognize a PDE signature in this sequence. However, linking to PLASMODB shows Pfam, PRINTS and SMART matches from InterPro for MAL13P1.118. All these matches refer to a PDE signature. Therefore, it can be assumed that the PRINTS search by pscan from EMBOSS is less sensitive than InterProScan.

From disorder predictions, Pfam domains, Superfamily domains and PDB matches, boundaries for functional structural domains can be identified. For the proteins investigated, low complexity regions were not useful for domain boundary identification.

SMID-BLAST predictions seem to be accurate but possibly not complete. The ligand scores calculated by SMID-BLAST are not reliable in all cases. SMID-BLAST is useful for functional characterization in cases where no Pfam, PRINTS and PROSITE matches are detected (MAL13P1.345).

Transmembrane predictions are indicative of protein localization within the cell and give insight into protein function. As with low complexity, coiled-coils, signal peptides and disorder, transmembrane regions make crystallization of proteins difficult. Disorder predictions never overlap with transmembrane helix predictions. Examples include all the predicted transmembrane proteins and can be selected via the query tool. Disordered regions are flexible and exposed to the surrounding solvent, therefore contain a high percentage of hydrophilic amino acids. In contrast, membrane spanning regions are hydrophobic and held intact by the membrane. For various proteins investigated, predicted disorder does not overlap with PDB matches. Examples include PF14_0020, MAL13P1.146 and

MAL13P1.159. Proteins with experimentally determined crystal structures had in general very little predicted disorder (PF08_0131, PFE0660c, MAL13P1.257). Membrane spanning regions also rarely overlap with PDB hits. Therefore, predicted transmembrane and disordered regions can be used to determine which proteins are suitable for experimental structure determination. In addition, proteins containing homologs in the PDB are of low priority for experimental structure determination. Proteins can further be prioritized according to indications of functional importance which are suggested by the functional family, protein interactions, exported proteins and interactions with small molecules.

The query tool allows for the selection of groups of proteins according to criteria set by a user. For example, a user can select proteins suitable for experimental structure determination by setting a minimum cut-off e-value for PDB matches, and setting cut-offs for the percentages of predicted disorder, coiled-coils and transmembrane content. Users are also able to select proteins for homology modeling by setting a maximum cut-off e-value and percent sequence identity for PDB matches. In addition, selections of proteins with a Pfam domain, PRINTS signature, PROSITE motif, export motif, a specific amount of transmembrane helices, interacting partners, small molecule binding sites or a crystal structure can be made in different combinations.

2.5. Conclusions

The structural annotation described here was designed to integrate a wide selection of structural feature predictions and apply it to the *P. falciparum*, *P. yoelii* and *P. vivax* proteomes. Predictions of secondary structure, disorder, transmembrane helices, coiled-coils, low complexity, signal peptides and small molecule interactions were included in the pipeline. Searches against the PDB, Superfamily, Pfam and PRINTS databases were performed and proteins were threaded through a library of SCOP folds. In addition, annotations from the literature include proteins exported to the RBC and protein-protein interactions. For the identification of domains and protein families, scanning against

integrated databases such as InterPro are more efficient and provide more coverage. Although PLASMODB contains many types of information also provided here, important features for structural characterization such as coiled-coils, disorder and small molecule binding site predictions are not included. In addition, threading is valuable supplement to searches against the PDB and Superfamily for template identification. Using the query tool presented here, specific groups of proteins can be selected according to user specifications. In PLASMODB, selections of proteins with a certain annotation can be made. However proteins with different combinations of annotations cannot be selected. Such selections are useful for the identification and prioritization of targets for experimental and computational structural studies and will be applied for this purpose in Chapter 3. Bioinformatics tools which could be added to the annotation pipeline, include automatic function prediction by ANN and a subcellular localization predictor. These annotations are not currently included in PLASMODB.

Chapter 3

Structural feature analysis of *Plasmodium* proteomes and putative target selection for structure determination

3.1. Introduction

Applications for proteome structural annotation include comparison of structural features between genomes (Liu and Rost, 2001; Ward *et al.*, 2004*b*; Rose *et al.*, 2005) and identification of targets for specific experimental and computational studies (Liu *et al.*, 2004). Structural genomics (SG) projects aim to solve the 3D structures of all the protein folds that exist by high-throughput X-ray crystallography and NMR. To realize this goal, protein structures containing novel folds must be determined. Therefore, the selection of targets has become a major priority of SG consortia such as the National Institute of Health (NIH) Protein Structure Initiative (PSI) centers. TargetDB (Chen *et al.*, 2004) is a database containing target information for projects of 19 SG centers over the world. The database keeps track of the progress of projects and contains information about the protein targets. Although SG projects aim to solve protein structures with novel folds, many proteins in TargetDB display high similarity to known folds. In addition, proteins that satisfy the selection criteria are often hypothetical. Therefore the amount of hypothetical proteins in the PDB has increased. These proteins are of limited value for structural-based drug design, since the protein function is unknown.

Previously, protein targets for the Northeast Structural Genomics Consortium (NESG) have been selected automatically by identifying those proteins which do not contain regions that complicate experimental structure determination (Liu *et al.*, 2004). Problem regions include transmembrane regions, signal peptides, proteins dominated by coiled-coils, low complexity and long regions without RSS. Proteins identified had to be shorter than 340 amino acids to avoid the problem of multi-domain proteins. Protein targets were selected based on the analysis of five eukaryotic model proteomes.

ProDomSG is a target selection service for structural genomics. From this server, proteins belonging to ProDom families can be selected according to similarity in the PDB. This selection procedure is only applicable to proteins containing domains within an existing protein family. Proteins without domains represented in ProDom are therefore not subjected to the selection process.

Malarial proteins are often longer than their homologs from other organisms and contain parasite-specific inserts. Various SG consortia attempt to solve one representative structure per protein family however, *Plasmodium* proteins often do not have enough similarity for assignment to a specific family. The Structural Genomics of Pathogenic Protozoa consortium (SGPP) focuses in particular on the determination of protein structures from *P. falciparum*. The PDB contains 210 *P. falciparum* protein structures of which at least a hundred display more than 90% identity. Therefore, malarial proteins in the PDB are redundant. To make structure determination efforts in the malaria research community more efficient, it is useful to construct a list of malarial protein targets for experimental structure determination.

In addition to experimental target determination, proteome structural annotation facilitates the identification of proteins suitable for homology modeling based on similarities in the PDB. The International Center for Genetic Engineering and Biotechnology (ICGEB) has constructed a pipeline for the automatic modeling of *P. falciparum* proteins

(`http://bioinfo.icgeb.res.in/codes/model.html`). However, the data from ICGEB has not been updated since 2005 and many new structures have become available in the meantime. New structures can offer better templates than those which have been used for the original modeling of the proteins. In addition, inserts have not been filtered out before these models were constructed, therefore the quality of these models are questionable. Identifying proteins for homology modeling and ranking them according to sequence similarity to structures in the PDB, will supply researchers with target lists for building high quality models. In addition, the target lists can be updated regularly as new information becomes available. Similar to the identification of proteins suitable for homology modeling, a list of proteins suitable for *in silico* docking studies can be constructed. Candidate proteins can be identified from the SMID-BLAST results and the annotation of proteins having experimentally determined structures.

Since many of the *P. falciparum* open reading frames are predicted proteins, targets for structural studies should fulfill certain criteria giving an indication of protein function. Protein interactions are an important step towards the identification of protein function for uncharacterized proteins (Eisenberg *et al.*, 2000). A protein interaction network has been constructed for the *P. falciparum* genome from yeast two-hybrid experiments (LaCount *et al.*, 2005). Computational modeling of the *P. falciparum* interactome using Bayesian networks, resulted in functional inferences for more than 2 000 uncharacterized proteins (Date and Stoeckert, 2006). Therefore, protein-protein interactions give an indication of the function of uncharacterized proteins. Using protein-protein interaction data together with Pfam domain identifications, protein targets for structural studies can be prioritized to help ensure that the predicted target proteins are not associated with pseudogenes.

Another application for structural annotation of proteomes is the comparison of structural features between proteomes. Comparison of features in *P. yoelli* and *P. falciparum* provides insight into modeling the disease in rodents. In addition, parasite specific inserts

can be studied by comparing *Plasmodium* species. Correlations between features can be detected by statistical analysis. Correlations are important for identifying features which can be used for predicting a related property. Previously, distributions of membrane, disordered and coiled-coil proteins were analysed for whole genomes. It is estimated that 20% of the proteins in eukaryotic species, including *H. sapiens,* have disordered regions longer than 50 amino acids long (Ward *et al.*, 2004*b*). Of the *P. falciparum* proteins encoded by chromosomes I and II, 25% contains predicted disordered regions longer than 50 amino acids long. Furthermore, 3% of these proteins are wholly disordered (Dunker *et al.*, 2000). No analysis of the whole proteome has been done so far. It is estimated that 20 to 30% of proteins in eukaryotic proteomes contains transmembrane helices (Krogh *et al.*, 2001). In another study, approximately 35% out of 10 000 *H. sapiens* proteins investigated, is transmembrane proteins (Wallin and von Heijne, 1998). Furthermore, approximately 8% of the *H. sapiens* proteome is predicted to contain coiled-coils, whereas 6% of the proteins in plant proteomes contains predicted coiled-coils (Rose *et al.*, 2005). Eukaryotic proteomes have been found to contain a larger portion (10%) of proteins, longer than a thousand amino acids, than prokaryotes (<5%). Protein composition is similar across species. Leucine, valine, serine and alanine are the most abundant amino acids and histidine, methionine, cysteine and tryptophan are the least abundant (Liu and Rost, 2001). Proteins interacting with other proteins, especially those interacting with 10 or more proteins, have been found to contain a high disordered content (Haynes *et al.*, 2006).

In this chapter, the occurrence of a series of structural features in the *P. falciparum* proteome are analysed and comparisons to *P. yoelii* and *P. vivax* are made. In addition, preliminary *P. falciparum* target lists for experimental and computational structural studies are proposed. The target lists are prioritized based on the suitability for a specific study and annotations indicating protein function. These target lists are stored in a relational database and can be updated when the structural information in the database is updated. The target lists will supply researchers with information about *P. falciparum*

Table 3.1: Information on the data used for structural feature analysis. WTSI refers to the Wellcome Trust Sanger Institute and SU refers to Stanford University.

PLASMODB FASTA file	Species & Strain	Sequencing	References
PfalciparumAnnotatedProteins_plasmoDB-5.0	*P. falciparum*, 3D7	WTSI, TIGR and SU	Gardner *et al.* (2002)
PvivaxAnnotatedProteins_plasmoDB-5.0	*P. vivax*, Salvador 1	Jane Carlton (TIGR)	Carlton (2003)
PyoeliiAnnotatedProteins_plasmoDB-5.0	*P. yoelii*, 17XNL	Jane Carlton (TIGR)	Carlton *et al.* (2002)

proteins least and most represented in the PDB. Furthermore, the annotations to each protein will be valuable for the design of experiments for structure determination.

3.2. Methods

3.2.1. Data sets

Three FASTA files were downloaded from PLASMODB version 5.0. Information about the sources and references of data in these files are summarised in Table 3.1.

The files obtained from PLASMODB contain amino acid sequences and annotations for the predicted proteins of each proteome. In this context, annotation includes terms such as 'hypothetical', 'conserved' and 'putative'.

3.2.2. Structural feature analysis for *P. falciparum*, *P.yoelli* and *P. vivax*

All proteome distributions regarding amino acid compositions, lengths, charge, IP, EC, number of transmembrane helices, disorder and low complexity were calculated by querying the database using Python scripts. Distribution graphs were constructed using the Pylab module which is a Python interface to Matplotlib.

3.2.3. Targets for homology modeling

Target classes for which high quality models can be constructed (HQ classes): An estimate of the proteins suitable for homology modeling was taken as the number of

proteins with BLAST/PDB hits better than a cut-off e-value of 1e-20. The list was then pruned by the removal of: proteins with experimentally determined structures; proteins with less than 30% identity to the PDB protein; and proteins which had PDB matches covering less than 70% of their sequence. In addition, proteins were prioritized according to annotations indicating protein function, i.e. interactions with other proteins and high similarity to functionally characterized protein families. Different priority classes were constructed in this way. Proteins within a class are of equal priority.

Target classes for which intermediate quality models can be constructed (IQ classes): These classes were constructed in the same way as described for the HQ classes with the exception of the cut-off thresholds. The PDB e-value cut-off was set at 1e-15, whereas the cut-off for sequence identity was changed to 25% and the sequence coverage cut-off was lowered to 60%. The proteins included in the HQ classes were removed.

3.2.4. Targets for *in silico* docking studies

Targets for *in silico* docking studies were identified by searching for proteins predicted to bind small molecules based on the SMID-BLAST search. Furthermore, an experimentally determined structure or a PDB match with a sequence similarity of 50% was required for the targets. The resulting set was prioritized based on protein-protein interactions, the presence of a transmembrane domain and the presence of export motifs.

3.2.5. Targets for experimental structure determination

A lack of similarity to structures in the PDB formed the basis for the putative identification of targets for X-ray crystallography. For this purpose, sequences were searched against the PDB without any cut-off e-value. Thus, the results reported all hits with e-values of up to 20 and higher. Proteins were classified by using the smallest e-value of all the hits for a specific query. Matches with e-values higher than 10 were assumed to be false and the query sequences were placed in the highest priority class (PC1) with

no PDB hits. Proteins were further classified according to e-value ranges as indicated in Figure 3.1. As the e-value range decreases, the priority becomes lower.

Predicted transmembrane helix, disorder, low complexity, signal peptide and coiled-coils content were employed for eliminating unsuitable targets. The identification process is represented grammatically in Figure 3.1. No proteins with e-values lower than 0.5 were selected. Predicted transmembrane and disordered regions are the least abundant in sequences for which experimentally determined structures exist. Furthermore, these two predicted features overlapped least with aligned sequences of known structure and the least with one another. Therefore, predicted transmembrane and disorder content formed the basis for the second round of elimination. Protein sequences containing more than 30% predicted disorder and transmembrane content were discarded. The third round of elimination was based on coiled-coil, signal peptide and low complexity content. Sequences from each PC containing more than 20% coiled-coils, signal peptides and low complexity all together, were excluded.

Finally, proteins in each PC were prioritized based on indications of function. Each PC was divided into two groups, one containing protein interacting proteins or proteins containing a Pfam domain (Group a), the other non-interacting proteins without Pfam domains (Group b). This prioritization resulted in twelve groups with PC1a forming the highest prioritization group.

3.3. Results

3.3.1. Statistical sequence analysis of the *P. falciparum* proteome

Property distribution

Figure 3.2 gives an overview of the distribution of properties for the *P. falciparum* genome. Twenty-seven percent of the proteins had BLAST/PDB hits with at least 25%

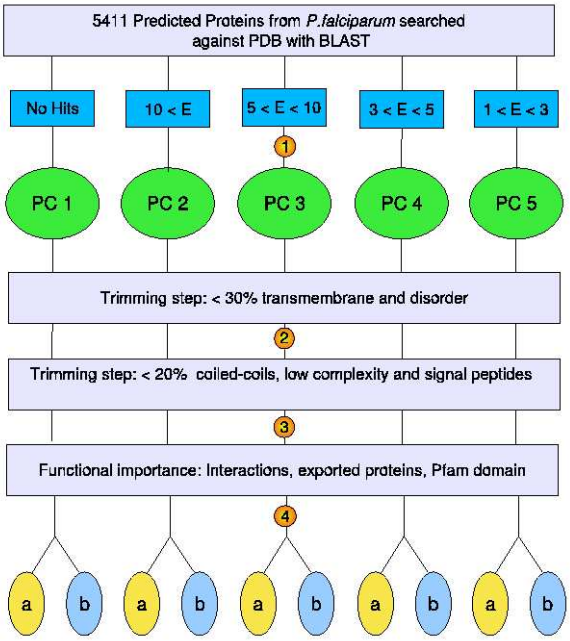


Figure 3.1: Flow diagram of the prioritization procedure followed for putative targets for experimental structure determination. In the first step, proteins containing PDB matches in ranges of e-values were selected to form six priority classes (PC). The sixth PC has an e-value range between 1 and 0.5 (not shown). In the second step, proteins with more than 30% transmembrane and disorder content in total were filtered out. Step 3 involved the elimination of proteins based on coiled-coil, low complexity and signal peptide content. The prioritization step (4) divided each PC into two groups. Proteins with a Pfam domain or protein interacting proteins were placed in group a and the remaining proteins were placed in group b.

identity to the hit. A comparison of the amount sequences and proportion of sequence covered by hits are shown in Figure 3.11. Thirty-two percent of the proteome had hits in Pfam below an e-value of 1e-15. At least 43% of all sequences had no hits with families in Pfam. Ten percent of the proteome had one or more predicted coiled-coil region, 5% are predicted to be transported out of the red blood cell based on the presence of the Pexel/VTS motifs, and about 30% of proteins were predicted to contain at least one transmembrane helix. It is notable that 37% of the proteins had a *Plasmodium* paralog. At least 22% of the proteome have been predicted to bind to small molecules by SMID-BLAST. Almost 15% of the proteome interact with other proteins according to the high-throughput yeast two hybrid experiments. Sixty percent of the proteins were predicted to contain at least 40% intrinsic disorder or no regular secondary structure. Finally, out of 2 462 proteins subjected to threading, 423 (8% of proteome) obtained sequence-structure alignments with Z-scores better than 3.95.

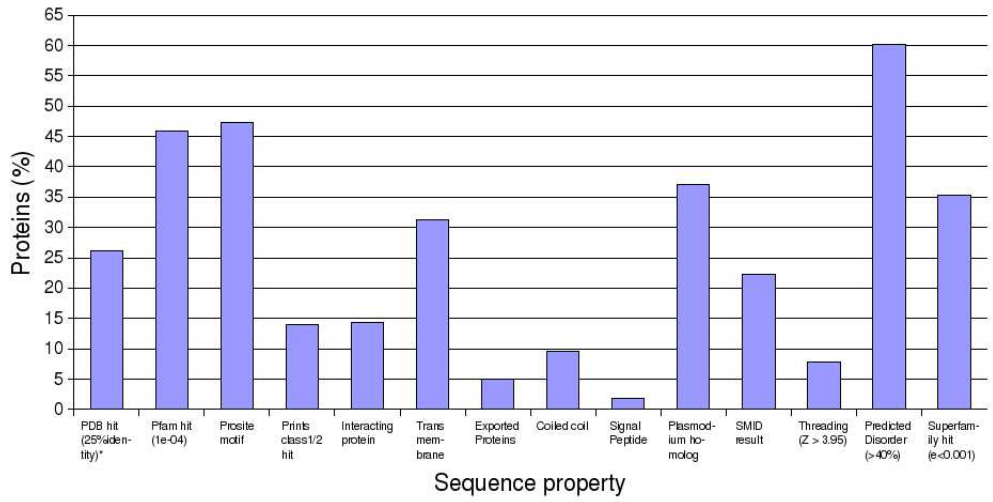


Figure 3.2: A summary of the predicted property distribution and coverage from different databases for the *P. falciparum* proteome. These properties were predicted by the tools described in Chapter 2. Cut-off values are indicated for database matches.

Sequence length distribution

Figure 3.3 shows the length distribution in *P. falciparum*. The mean of the lengths is 740 amino acids with a standard deviation of 808 amino acids, indicating the high variation in protein lengths.

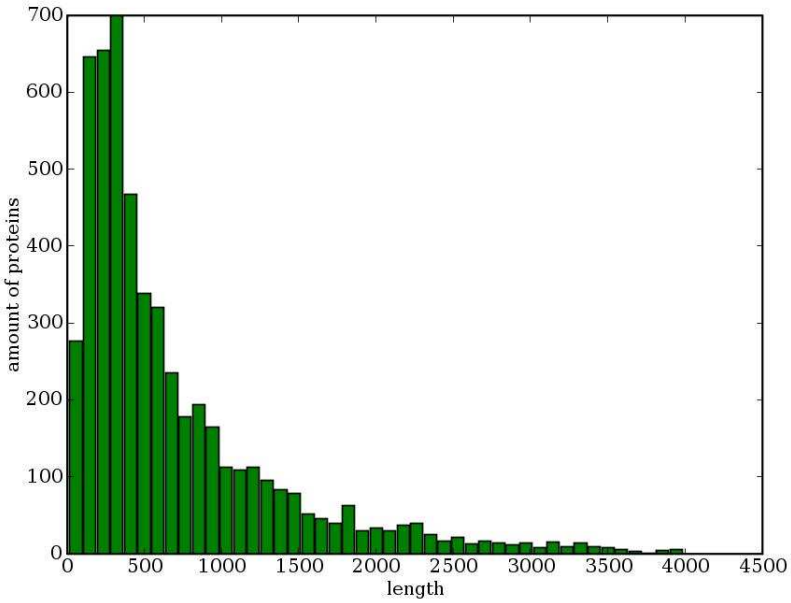


Figure 3.3: *P. falciparum* protein length distribution.

Composition

Figure 3.4 shows the relative distribution of amino acids in the *P. falciparum* proteome. The most abundant amino acid is asparagine and the least abundant amino acid is tryptophan. Figure 3.5 shows the relative percentages of negatively charged, positively charged, polar uncharged, aliphatic, aromatic and the remaining non-polar residues in the *P. falciparum* proteome. Additionally, 43.6% of the proteome consists of small residues (A, C, D, G, N, P, S, T, V) and tiny residues (A, C, G, S, T) make up 18.2% of the proteome.

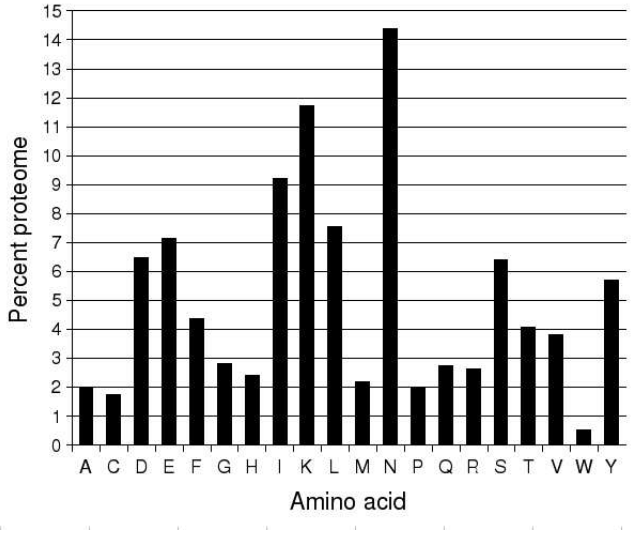


Figure 3.4: Amino acid composition of the *P. falciparum* proteome.

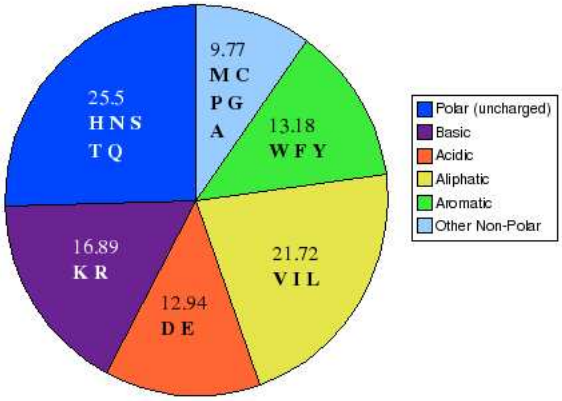


Figure 3.5: Amino acid type distribution in *P. falciparum*.

Protein charge, IP, and the extinction coefficient distributions

The charge distribution is shown in Figure 3.6. The curve is more or less symmetrically bell shaped. The mean charge is 15.4 and the standard deviation is 42.8. There are 60 proteins with no overall charge. Figure 3.7 shows the distribution of IPs of *P. falciparum* proteins as calculated by EMBOSS Pepstats. The mean IP is 8.04 and the standard deviation is 1.93.

Figure 3.8 shows the extinction coefficient distribution of the *P. falciparum* proteins. The mean EC is 0.91 with a standard deviation of 0.39.

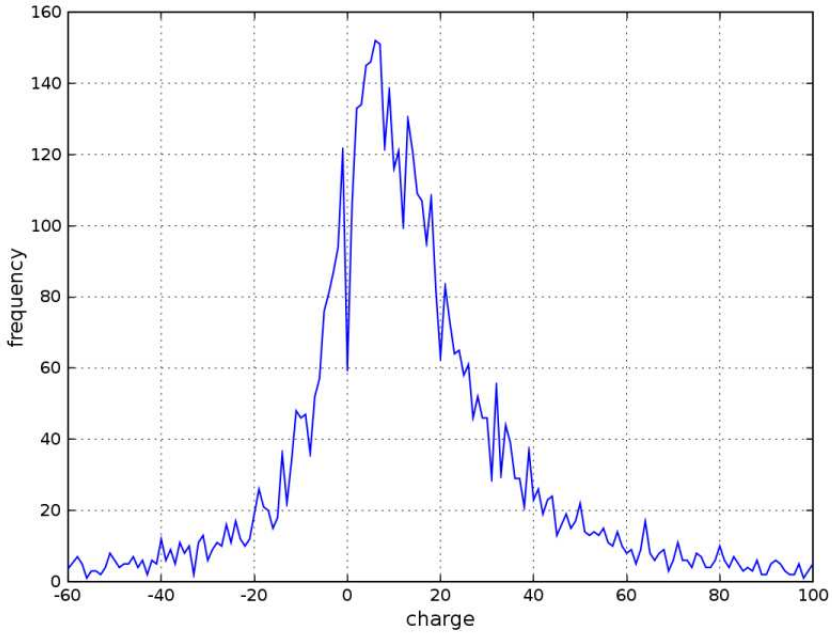


Figure 3.6: Charge distribution for the *P. falciparum* proteome.

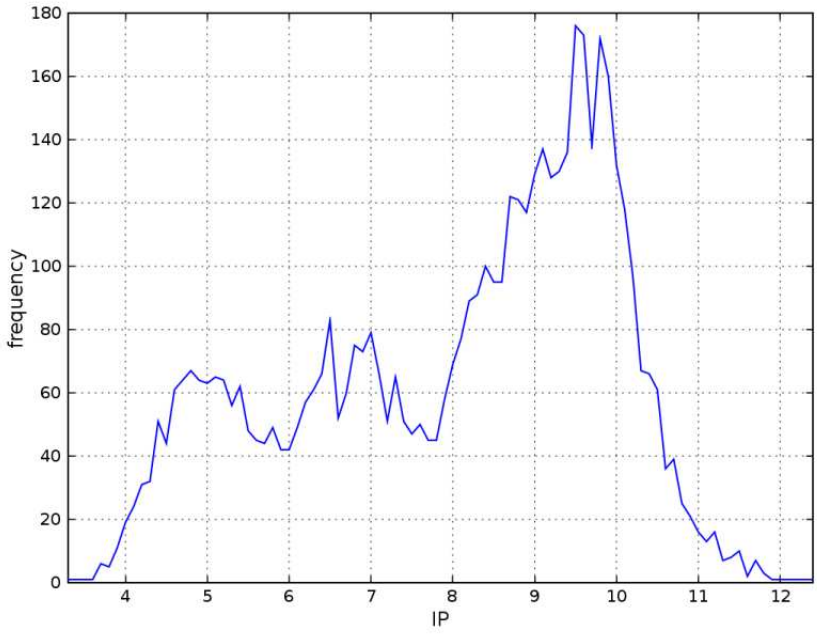


Figure 3.7: Iso-electric Point distribution of the *P. falciparum* proteome.

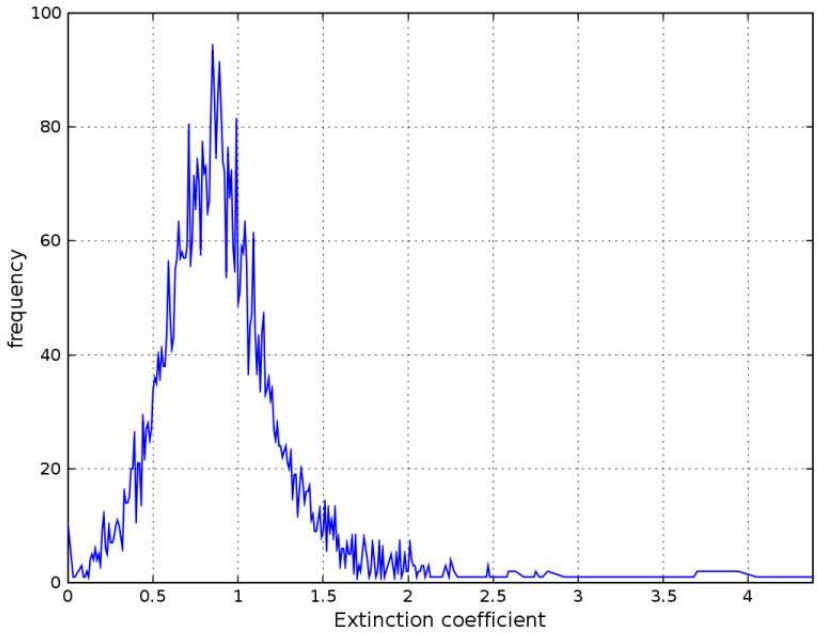


Figure 3.8: Extinction coefficient of the *P. falciparum* proteome.

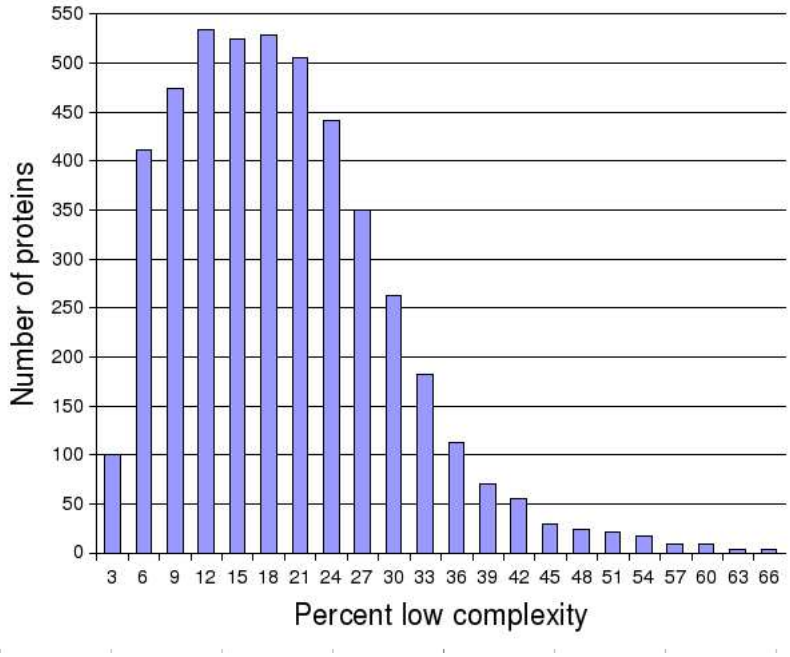


Figure 3.9: Distribution of low complexity in the *P. falciparum* proteome as calculated by SEG.

Low complexity

Figure 3.9 shows the distribution of low complexity regions in *P. falciparum*. The number of sequences without low complexity regions is 725. The mean percentage low complexity per sequence is 16% with a standard deviation of 9%.

Family assignment

Figure 3.10 shows the contribution of the searches against the different family databases Pfam, PRINTS, PROSITE and Superfamily. PRINTS class 1 refers to hits where all the elements in the fingerprint are present in the correct order in the query sequence. PRINTS class 2 refers to hits where all the elements in the fingerprint are present in the query sequence but in a different order as specified by the fingerprint. These two classes of hits are most reliable for assigning a functional family to the query sequence. Class 3 refers to hits where not all the elements in a fingerprint are present in the query sequence, but those that are present are in the same order as in the fingerprint. Class 4 describes similar hits as class 3 except that the elements are not in order.

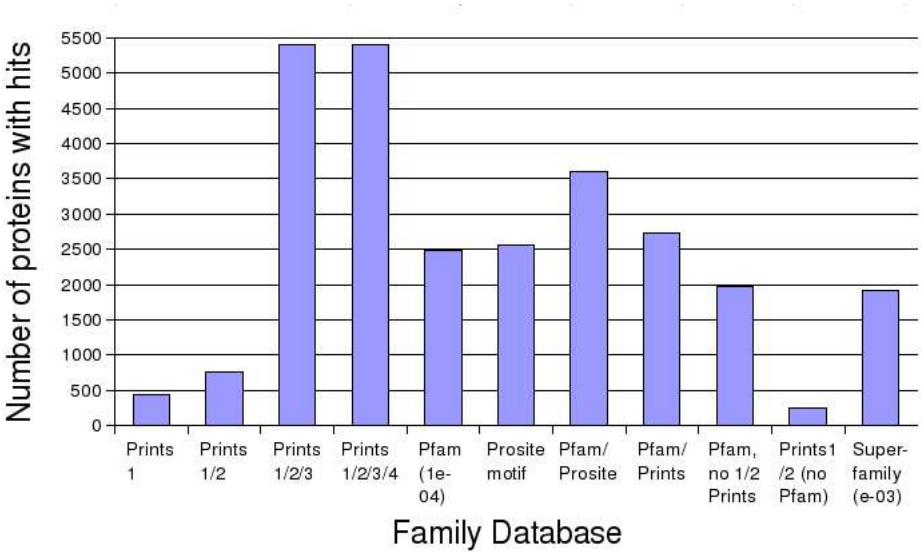


Figure 3.10: Relative contributions of different family databases to family assignment. Prints 1,2,3,4 refer to Prints class1, class2, class3 and class4 hits respectively. The Pfam cut-off e-value was set at 1e-04.

There were 1 987 sequences with hits in Pfam with an e-value smaller than 1e-10. An additional 332 sequences had PRINTS class 1 or class 2 hits. Thus, a total of 2 319 sequences or 43% of the proteome could be assigned to functional families making use of the Pfam and PRINTS databases. A large portion of proteins contained frequently occurring PROSITE hits, therefore a reliable estimate of the amount of proteins which could be assigned to a family making use of PROSITE, could not be made. Almost 200 proteins had Superfamily hits better than 1e-03. Of these, 650 sequences did not have Pfam or PRINTS matches.

Fold assignment

Figure 3.11 shows the amount of sequences having BLAST-PDB hits with e-values lower than 1e-10 and the the respective percent residues aligned for the sequences. Ten percent of the proteins had two-thirds of their sequence covered by a PDB match with an e-value better than 1e-05. Almost 20% of the sequences had at least one-third sequence covered by a PDB protein. An additional 413 sequences had Superfamily hits with a score of 100 or more. Therefore, an estimate of proteins which could be assigned to

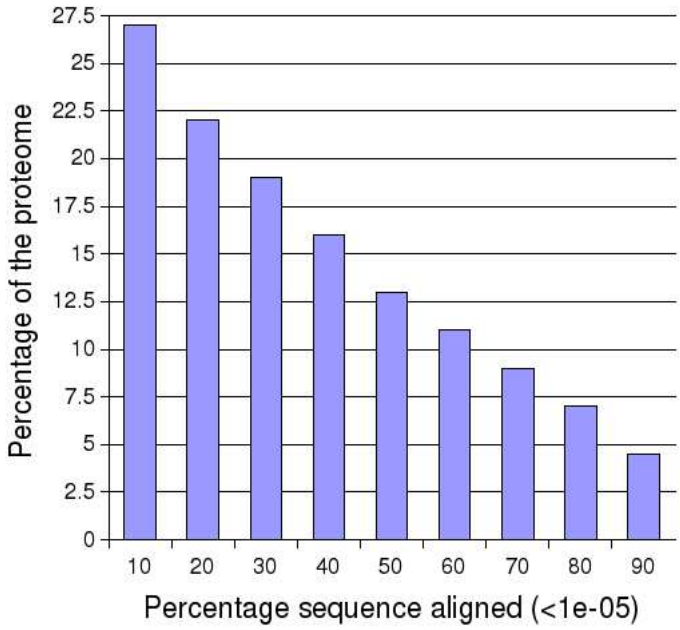


Figure 3.11: The sequence coverage of BLAST/PDB hits (e-value < 1e-05) for proteins from *P. falciparum*.

an existing fold is 1 224 or 23%. Of the proteins which were subjected to threading, 423 had alignments with Z-scores better than 3.95. Of these, about 100 did not have BLAST-PDB matches with e-values smaller than 0.5, which covered more than 30% of the query sequence.

Transmembrane proteins

Figure 3.12 shows the percentages of *P. falciparum* transmembrane proteins with different amounts of transmembrane helices. As with other genomes, the most abundant transmembrane proteins contain only one transmembrane helix (Wallin and von Heijne, 1998). The amount of transmembrane proteins decrease as the amount of membrane spanning helices increase, with the exception of 6-tm and 11-tm proteins which are slightly more than the portion of 5-tm and 10-tm proteins, respectively.

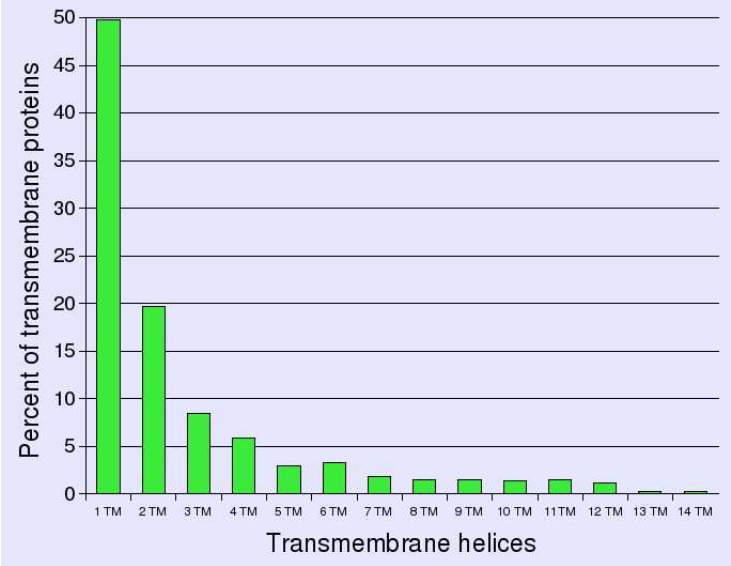


Figure 3.12: Transmembrane helix distribution for *P. falciparum* proteome.

Intrinsic disorder in protein-interacting proteins

The mean percentage disorder in interacting sequences is 61%, while the mean percentage disorder in non-interacting proteins is 44% and the overall mean percentage disorder for all sequences is 48%. As expected, interacting proteins contain a higher intrinsic disorder content than non-interacting proteins. Because disorder in a protein makes it more flexible, it was expected that the disorder content would increase with the number of interacting partners. Figure 3.14 shows how the percent of disorder in *P. falciparum* proteins vary with the amount of interacting partners. For proteins interacting with only one other protein the predicted disorder varies from 4% to 100%. The majority of interacting proteins interact with less than 10 other proteins. As the amount of proteins decreases with an increasing number of interacting partners, the range of variation in disorder in the proteins also decreases, as expected. The ranges tend to span higher percentages of disorder as the amount of interacting partners increase. Figure 3.15 shows the disorder distributions for proteins interacting with one other protein, and proteins interacting with 10 or more other proteins.

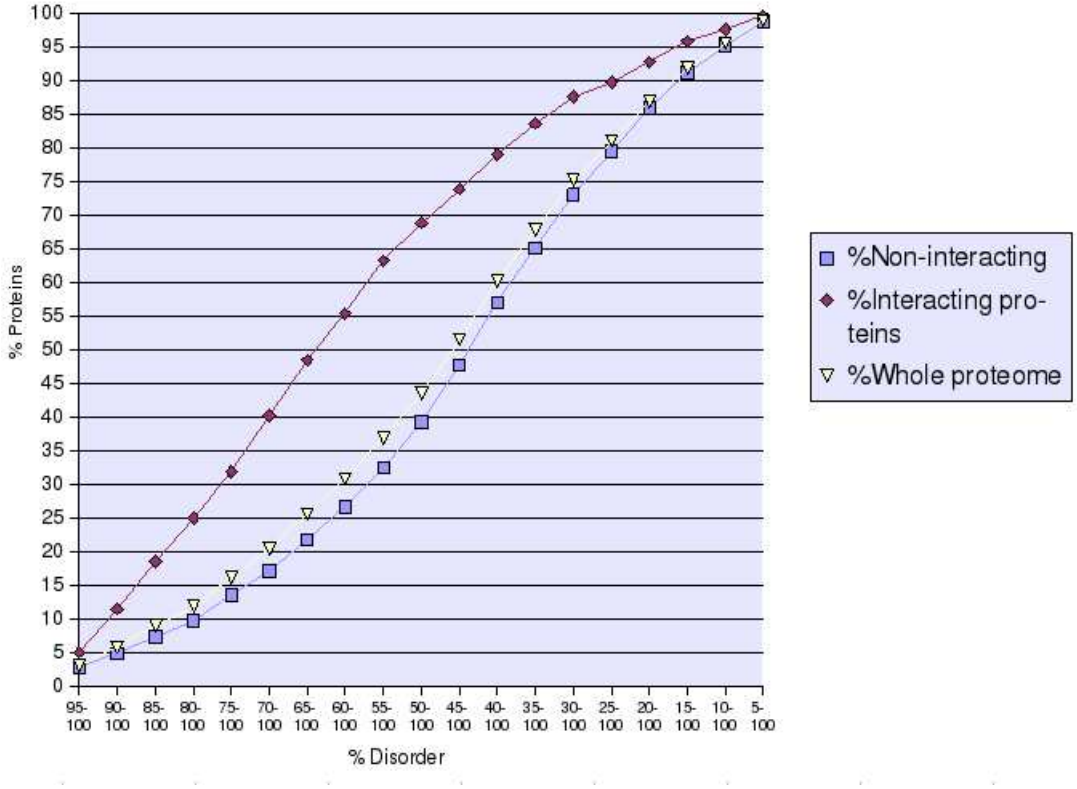


Figure 3.13: Correlation between disorder and interacting proteins in *P. falciparum*.

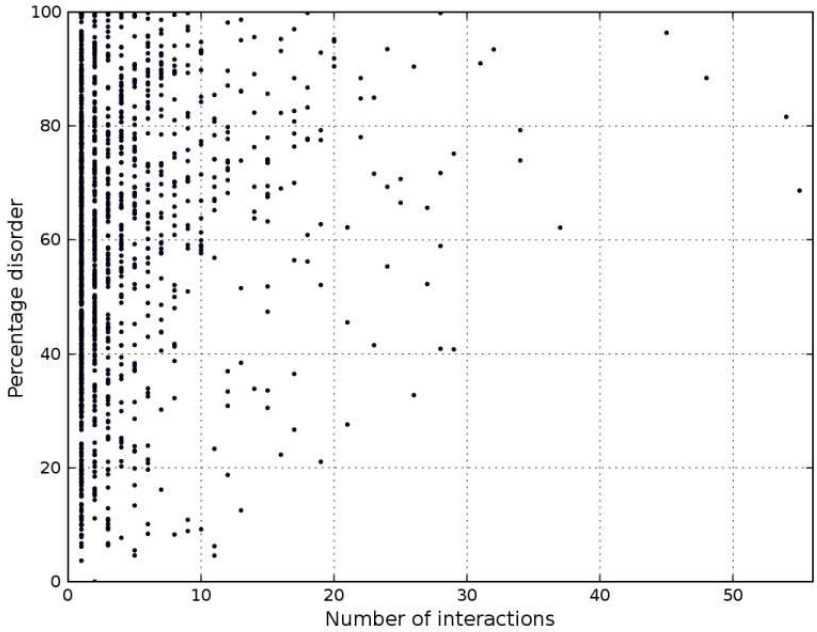
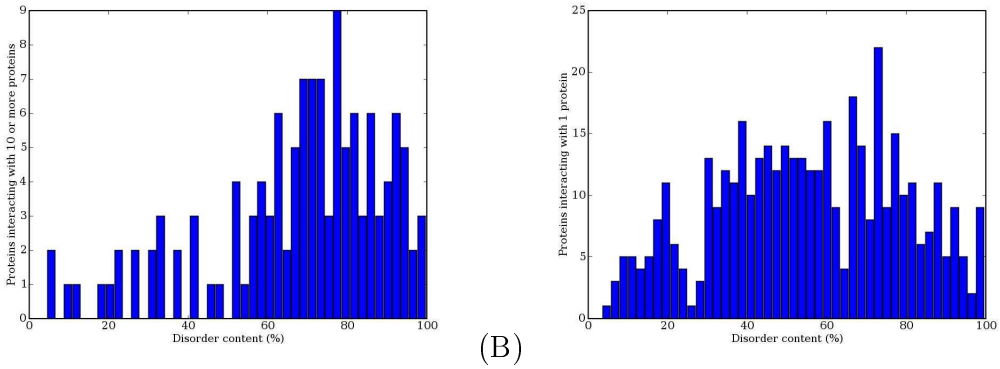


Figure 3.14: Distribution of predicted disorder in interacting proteins in *P. falciparum*.



(A) (B)

Figure 3.15: (A) Disorder content in proteins interacting with 10 or more proteins. (B) Disorder content in proteins interacting with only one other protein.

3.3.2. Species comparison

Figure 3.16 shows the distribution of lengths of amino acid sequences in *P. vivax* and *P. yoelii*. The proteins from the *P. falciparum* distribution has a longer tail than the other two species, and *P. yoelii* has a more symmetrical length distribution than the other species. The mean length for *P. vivax* is 630 with a standard deviation of 576 amino acids and the mean length for *P. yoelii* is 420 with a standard deviation of 450. The proteins in *P. yoelii* vary less in length than in the other two species, with *P. falciparum* showing the most variation.

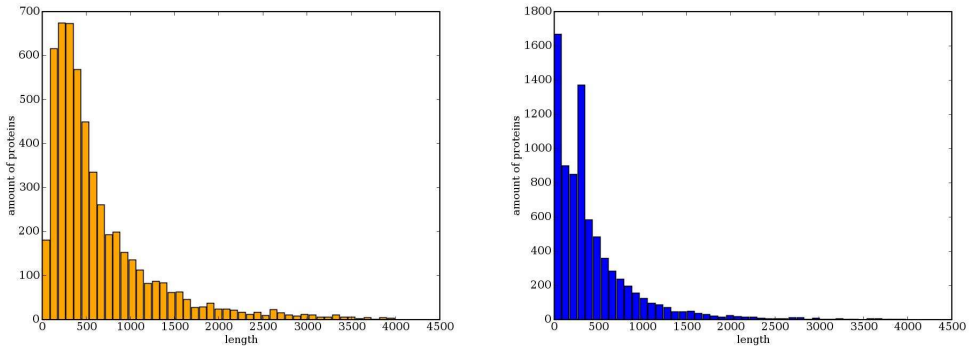


Figure 3.16: *P. vivax* and *P. yoelii* protein length distributions.

Figure 3.17 shows the amino acid distributions over the sequences for the three *Plasmodium* species. Asparagine is the most abundant amino acid in *P. falciparum* and *P. yoelii*, and Lysine in *P. vivax*. Although lysine is the most abundant amino acid in *P.*

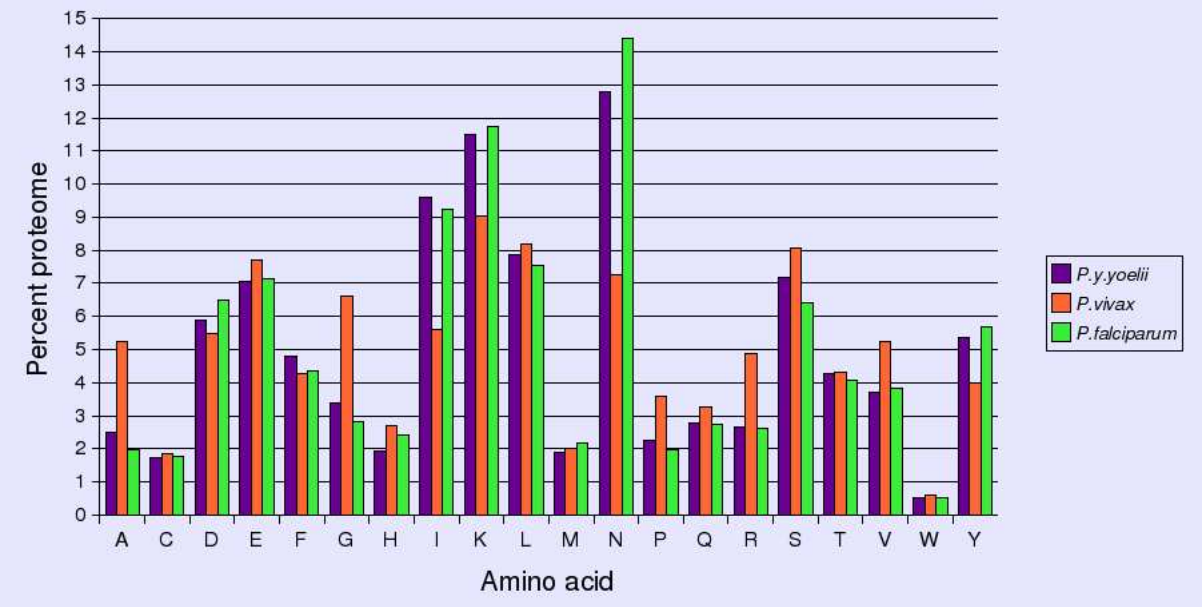


Figure 3.17: Amino acid distributions of *P. vivax*, *P. yoelii*, and *P. falciparum*.

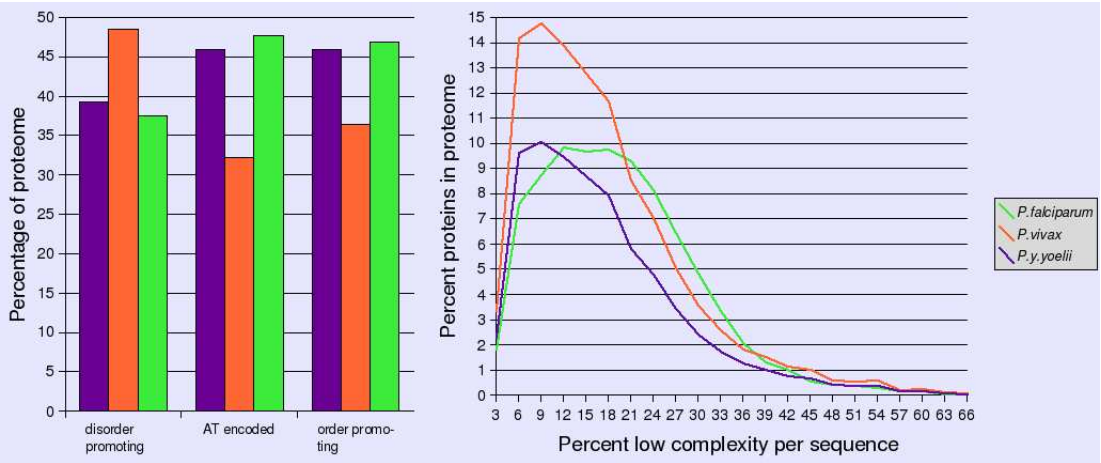


Figure 3.18: Distribution of low complexity in *P. falciparum*, *P. vivax* and *P. yoelii*.

vivax, it should be noted that Lysine is less abundant in *P. vivax* (9%) than in the other two species (11.5%). *P. vivax* contains on average twice as many alanine and glycine as the other two species. Overall, 26% of residues in *P. vivax* is tiny, in comparison to the 18% and 19% tiny residues contained within *P. falciparum* and *P. yoelii*, respectively.

General statistics for the proteomes of *P. falciparum*, *P. vivax* and *P. yoelii* are summarised in Table 3.2. Low complexity comparisons are represented by Figure *3.18*. *P.*

Table 3.2: Comparison of general sequence properties of three *Plasmodium* species.

Property	***P. falciparum***	***P. vivax***	***P. yoelii***
No. of Proteins	5 411	5 352	7 864
Mean length	740	609	433
Mean charge	15.4	14.4	7.7
Mean IP	8.04	7.97	8.09
Mean EC	0.91	0.82	0.94
Most abundant aa	Asn	Lys	Asn

vivax contains more proteins with small percentages of low complexity. Although *P. yoelii* and *P. falciparum* contain the same amount of proteins with predicted low complexity regions, the proportion of *P. yoelii* proteins is much lower than for *P. falciparum*. Figure *3.18* also displays comparisons between disorder promoting and order promoting amino acids. *P. yoelii* and *P. falciparum* have similar proportions of disorder and order promoting amino acids, whereas *P. vivax* has proportionally more disorder favouring amino acids and less order promoting amino acids. Figure 3.19 shows the differences in coverage from various databases as well as differences in predicted properties between the species. In general, *P. falciparum* is better covered in the databases than *P. vivax* and *P. yoelii*, while the latter has the least coverage. *P. yoelii* also has the least coiled-coil regions. The low complexity is represented as the total percentage of low complexity regions in all the sequences. The average percent low complexity per sequence is 16% in *P. falciparum*, 10% in *P. vivax*, and 12% in *P. yoelii*. No low complexity is predicted for 27%, 19% and 13% percent of the sequences in *P. yoelii*, *P. vivax* and *P. falciparum*, respectively. *P. yoelii* and *P. falciparum* have an equal portion of transmembrane proteins, while *P. vivax* has less predicted transmembrane proteins. *P. falciparum* has more 2-tm, 3-tm, 4-tm, 6-tm and 9-tm proteins than the other two species. *P. vivax* has slightly more 8-tm proteins than *P. yoelii* and *P. falciparum* and *P. yoelii* has the most 1-tm proteins.

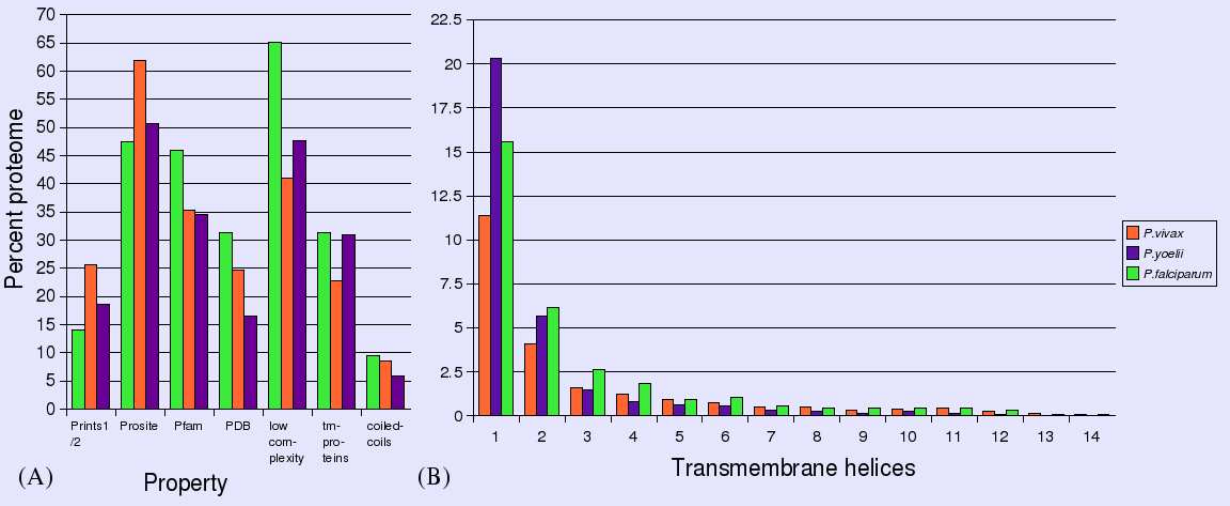


Figure 3.19: (A) Comparison of the predicted properties and database coverage of the proteomes. (B) Comparison of the transmembrane protein distributions.

3.3.3. Target identification in *P. falciparum*

Identification of proteins suitable for homology modeling

Tables containing candidates suitable for homology modeling can be viewed through the web interface (`http://deepthought.bi.up.ac.za:8080/Annotation/modeling`). These proteins contain PDB matches with e-values less than 1e-20 and which have more than 70% of their sequence covered by the PDB match. The cut-off sequence identity was set at 25%. Therefore, these tables contain proteins for which high quality models can possibly be obtained through automatic model building. Separate tables contain interacting proteins, proteins with Pfam domains and uncharacterized proteins (`http://deepthought.bi.up.ac.za:8080/Annotation/modeling2`).

Table 3.4 is a subset of proteins which had e-values smaller than 1e-25, more than 70% of their sequence covered by the hit and more than 40% identity with the PDB sequence. Many of these proteins are DNA and RNA interacting proteins, therefore involved in gene regulation. Proteins involved in gene regulation are probably more conserved across species and therefore these proteins have the best hits in the PDB. In addition, many researchers have focused on the determination of structures for proteins involved in gene

regulation because these proteins are important drug targets.

Proteins suitable for in silico docking studies

Proteins possibly suitable for *in silico* docking studies can be accessed through the web interface (`http://deepthought.bi.up.ac.za:8080/Annotation/docking`). These proteins were selected based on the presence of predicted small molecule binding sites and the availability of a 3D structure. Interacting proteins were separated from non-interacting proteins. Proteins with a determined crystal structure are also listed separately. The small molecules predicted to bind to the proteins could be manually examined for drug-like properties, inhibitory properties, etc. In addition, the proteins could be compared to human proteins to identify differences in binding sites. Malarial proteins which differ structurally from their human homologs can then be identified as putative drug targets. All of these proteins have an experimentally determined structure or very high similarity to an experimentally determined structure. Table 3.3 is a subset of proteins from the protein-interacting proteins which have functional annotation. These proteins obtained the highest binding scores in the SMID-BLAST results and they are predicted to bind to a variety of small molecules.

Table 3.3: A subset of proteins identified for *in silico* docking studies.

PlasmoDB ID	Description	No of Molecules	References
PFC0975c	PFCYP19, cyclophilin, peptidyl-prolyl cis-trans isomerase	15	PDB, CDD, BIND
PF14_0223	cyclophylin, putative	11	PDB, CDD, BIND
PF11_0161	falcipain-2 precursor, putative	12	PDB, CDD, BIND
PF14_0425	fructose-bisphosphate aldolase	4	PDB, CDD, BIND
PF10_0232	hypothetical protein	11	PDB, CDD, BIND
PF08_0126	DNA repair protein, putative	10	PDB, CDD, BIND
PFL1725w	ATP synthase beta chain, mitochondrial precursor	10	PDB, CDD, BIND
PFD0230c	protease, putative	15	PDB, CDD, BIND
PF14_0223	calmodulin	11	PDB, CDD, BIND

Table 3.4: Candidate proteins identified for homology modeling. The proteins are sorted according to the percent sequence identity. % Seq refers to the percentage of the query sequence which aligned to the PDB match.

	PlasmoDB ID	PDB hit	Description	E-value	% Seq	%ID
1	PFF0865w	2F8N	histone h3	2e-58	99	81
2	PF14_0124	1NLV/1D4X	actin II	1e-172	99	75
3	PFE1005w	1S1H|D	40S ribosomal subunit protein S9*	5e-60	83	70
4	PFA0345w	2GGM	centrin, putative	2e-62	99	69
5	PF10_0155	2AKZ	enolase	1e-163	98	69
6	PFEO845c	1S1I|B	60S ribosomal subunit protein L8*	3e-91	96	62
7	MAL8P1.69	2C1N/1QJA	14-3-3 protein homologue*	8e-80	93	60
8	PF14_0655	1FUU	RNA helicase-1*	1e-128	97	60
9	PF10_0272	1S1I|C	ribosomal protein L3*	1e-135	99	59
10	PFI1090w	2HJ2	s-adenosylmethionine synthetase*	1e-121	97	54
11	PFC0255c	1JAT	ubiquitin-conjugating enzyme E2*	1e-37	94	52
12	PF14_0443	2GGM	centrin*	5e-41	91	52
13	PF07_0117	1Q8K	eukaryotic TIF 2 alpha subunit*	1e-72	81	52
14	PFL2060c	1LV0/1GND	rabGDI protein	1e-127	97	51
15	PF14_0492	1M63/1MF8	protein phosphatase 2b regulatory subunit*	4e-44	90	51
16	PF10_0038	1S1H|J	ribosomal protein S20e*	3e-24	84	49
17	PFE0350c	1s1I|D	60S ribosomal subunit protein L4/L1*	1e-97	90	49
18	PFF0530w	1TKC/1GPU	transketolase*	1e-176	98	49
19	PF10_0086	2AK2	adenylate kinase*	7e-56	81	49
20	PFL2095w	2IF1	Translation intiation factor, SUI1*	8e-26	96	49
21	PF11_0351	1YUW	heat shock protein hsp70 homologue	1e-147	82	49
22	PFA0400c	2GPL/1Z7Q	beta3 proteasome subunit*	2e-50	100	47
23	PFE0285c	1L2N	ubiquitin-like protein*	3e-18	94	46
24	PF08_0059	1RZY	protein kinase C inhibitor-like protein*	6e-28	99	46
25	PF13_0268	1S1I|N	ribosomal protein L17 *	2e-39	82	46
26	MAL13P1.337	1P22	Skp1 family protein*	2e-32	95	45
27	PF11_0245	1R5O	translation EF-1, subunit alpha*	1e-109	83	45
28	PF14_0391	1S1I|A	ribosomal protein L1*	2e-53	99	44
29	MAL13P1.344	1YQT	RNAse L inhibitor protein*	1e-123	85	43
30	PF07_0088	1IQV	40Sribosomal protein S5*	4e-35	96	41

* These proteins have putative descriptions

Experimental structure determination

Twelve lists of proteins resulted after the prioritization procedure (Figure 3.1) was followed. These lists can be obtained through the web interface at `http://deepthought.bi.up.ac.za:8080/Annotation/experimental1` (proteins with a Pfam domain) and `http://deepthought.bi.up.ac.za:8080/Annotation/experimental2` (proteins without a Pfam domain). The number of proteins in each list after every step is given in Table 3.6. The numbers of functionally characterized and uncharacterized proteins are also shown (groups a and b). Table 3.7 describes the properties of proteins in PC1a, PC2a, PC3a and PC4a and Table 3.8 shows the Pfam domains within the target proteins.

Table 3.6: The number of targets identified for each priority class after every elimination step. Tm+disorder refers to the amount of proteins in each group after the transmembrane and disorder filtering step. CC+LC+SP refers to the proteins in each group after coiled-coils, low complexity and signal peptide filtering step. Group a indicates proteins containing a Pfam functional domain.

Priority Class (PC)	PDB E-value range	No. of proteins	Tm+disorder	CC+LC+SP	a	b
PC1	No PDB matches	139	11	8	1	7
PC2	E-value > 10	174	15	11	3	8
PC3	10 >= E-value > 5	352	34	31	9	22
PC4	5 >= E-value > 3	332	35	19	5	14
PC5	3 >= E-value > 1	810	88	51	12	39
PC6	1>= E-value > 0.5	539	60	58	10	48

The proteins in Table 3.7 do not contain any signal peptides. Five proteins contain one predicted transmembrane helix and another protein contains three predicted transmembrane helices. The disorder and transmembrane predictions in these proteins do not overlap. In contrast, low complexity predictions frequently overlap with transmembrane and disorder predictions. In addition, coiled-coils often overlap with disorder predictions. One protein does not have a functional annotation although it contains a Pfam domain. The matching Pfam family is described as a 'Domain of unknown function'. Another observation is that many of the proteins are interacting with DNA. For most of the proteins, the predicted disorder regions are small and separated and have low confidence prediction

Table 3.7: High priority putative protein targets for experimental structure determination. PC indicates the priority class as explained in the text. E-value refers to the smallest e-value obtained in a BLAST-PDB search. Tm%, Dis% and LC% refers to the predicted transmembrane, disorder and low complexity content as a percentage of the sequence, respectively. Int indicates the number of protein interactions with the target protein. None of these proteins contained coiled-coils or signal peptides.

PLASMODB ID	Description	PC	E-value	Tm %	Dis%	LC %	Int
MAL8P1.67	conserved hypothetical protein	PC1a	no hit	0	28.6	10.3	0
PFC0520w	26S proteasome regulatory subunit S14*	PC2a	19	0	29.6	8.22	0
PF10_0144	Prohibitin *	PC2a	11	0	23	4.6	0
PFL0620c	Choline transporter	PC3a	8.4	17.3	9.9	0	1
PFL1655c	Hypothetical protein	PC3a	5.3	0	26.3	10	0
PFL0825c	Hypothetical protein	PC3a	5.2	0	20.4	14.7	0
MAL13P1.190	proteasome regulatory component *	PC3a	5.5	0	25	12.5	0
PFB0880w	Hypothetical protein, conserved	PC3a	10	0	17.5	13.3	1
PFB0260w	proteasome 26S regulatory subunit *	PC3a	6.5	0	28.2	10.6	2
PF10_0062	hypothetical protein	PC3a	8.3	0	28	6.9	0
PFF0615c	membrane protein pf12 precursor	PC3a	5.9	6.3	16.1	9.8	0
PFA0225w	LytB protein	PC4a	4.5	3.5	23.9	6.5	0
PF08_0050	hypothetical protein	PC4a	4.3	2.9	23.5	7.7	0
PF13_0224	60S ribosomal subunit protein L18 *	PC4a	4	0	24.5	12.5	0
PFE0395c	hypothetical protein	PC4a	3.5	5.4	9.2	8.3	0
PFE1115c	methyltransferase *	PC4a	4.3	0	28.3	11.8	0

*These proteins have putative descriptions

Table 3.8: Pfam domains contained within the proteins in Table 3.7.

PLASMODB ID	Pfam Domain (e-value < 1e-04)
MAL8P1.67	Domain of Unknown Function
PFC0520w	SAC3_GANP
PF14_0671	Uncharacterized protein family
PF10_0144	Band_7
PFL0620c	Acyltransferase
PFL1655c	DNA Polymerase alpha/epsilon subunit
PFL0825c	DNA Topoisomerase
MAL13P1.190	Proteasome regulatory subunit
PFB0880w	Saccharopine dehydrogenase
PFB0260w	Proteasome/cyclosome repeat
PF10_0062	NOT2 / NOT3 / NOT5 family
PFF0615c	Sexual stage antigen s48/45 domain
PFA0225w	LytB protein
PF08_0050	MAC/Perforin domain

Table 3.9: General sequence features of identified targets and proteins with crystal structures. LC and CC refer to the average percentages of predicted low complexity and coiled-coils per sequence. IEB refer to the average improbability of expression of inclusion bodies. Unfavourable is the sum of percentages of predicted disorder, low complexity, transmembrane helix and coiled-coils content.

Proteins	Length	IP	EC	Disorder	Tm helices	IEB	LC	CC	Unfavourable
Putative targets	354	7.4	1.1	19 %	1.5 %	0.73	8.4 %	0.15 %	29.05 %
Crystal structures	424	8.3	1.0	24.7 %	3.7 %	0.74	5.8 %	0.47 %	34.67 %

values (<6.5). Four proteins in Table 3.7 are interacting with other *P. falciparum* proteins of which one is hypothetical and does not contain Pfam domains.

Comparison of identified experimental targets to proteins with crystal structures

Table 3.9 shows a comparison of the averages of general sequence features between the putatively identified targets (PC1 to PC6) and *P. falciparum* proteins with crystal structures. The probability of expression in inclusion bodies is a measure of the proteins' solubility. For example, proteins can be expressed in *E. coli* either in a soluble form or in an insoluble form in inclusion bodies. The solubility of the protein does not only depend on the sequence but it can be valuable for comparison between proteins with a solved experimental structure and putatively identified targets.

3.4. Discussion

3.4.1. Disorder and interactions in *P. falciparum*

The predicted disorder content in interacting proteins is higher than that for non-interacting proteins. Previous findings have shown that the disorder content in human proteins interacting with ten or more proteins, is higher than that of proteins interacting with one protein. For *P. falciparum*, the disorder content of proteins interacting with one protein is more or less evenly distributed. For proteins interacting with ten or more proteins, the distribution is more concentrated in the 60-100% range.

3.4.2. Comparisons between predicted features in the *Plasmodium* proteomes

General statistics

Statistical calculations such as IP, EC and charge are comparable for all three species. *P. falciparum* have on average the longest proteins. *P. yoelii* and *P. vivax* proteins are on average 300 and 100 residues shorter, respectively. *P. yoelii* and *P. falciparum* have similar amino acid distributions. *P. vivax* has an amino acid distribution which more closely resembles those of other eukaryotes with larger percentages of Ala, Gly, Ser, Arg and Val and smaller percentages of Asn, Lys, Ile and Tyr than *P. falciparum* and *P. yoelii*. These findings imply that *P. vivax* and *P. yoelii* have less inserts than *P. falciparum* and comparisons between these species can therefore be used to identify inserts in *P. falciparum*, without filtering out functional regions specific to *Plasmodium*. The *P. yoelii* and *P. vivax* orthologs of *P. falciparum* proteins, can also act as intermediates for identifying other similar proteins.

Low complexity and disorder

Low complexity regions are non-globular and similar to disordered regions in a protein. Therefore, it is expected that a proteome with more disorder-promoting amino acids should also contain more low complexity regions. Although *P. falciparum* contains more predicted disorder and low complexity per sequence than *P. vivax* and *P. yoelii*, *P. falciparum* contains the least disorder-promoting and most order-promoting amino acids. One explanation for this anomaly is that low complexity and disorder predictions make use of a window size, thus taking neighbouring residues into consideration. *P. vivax* contains the most disorder-promoting amino acids. *P. yoelii* contains the least predicted average low complexity and disorder percentages per protein. The portion of amino acid chains in *P. yoelii* containing disordered regions longer than 50 amino acids (33%) is considerably less than the corresponding portions in *P. falciparum* (67%) and *P. vivax* (60%). One possibility is that *P. yoelii* contains more or less the same amount of proteins

than *P. vivax* and *P. falciparum* but the large number of 'false' proteins (2 000) causes the relative percentage of disordered proteins to be less. The percentages of proteins in *P. falciparum* and *P. vivax* containing predicted disordered regions longer than 50 amino acids are almost double than previously predicted for other eukaryotes, including *H. sapiens*. Different disorder prediction methods were used for these predictions. Therefore, the discrepancies could be due to the VSL2 predictor over-predicting disorder. One possible explanation for a higher disorder content in *Plasmodium* proteins, lies in the parasite's adaptations to avoid the immune system of the host. If the disorder content in *Plasmodium* proteins is really higher than for other eukaryotes, mutations probably do not influence the parasite's proteins as much as the host's. While a high mutation rate can help the parasite to evade the immune system and gain resistance to drugs, the loss in protein function is probably not significant due the flexibility of the protein.

Transmembrane protein distribution and species comparison

P. yoelii and *P. falciparum* have an equal portion of transmembrane proteins, while *P. vivax* has less transmembrane proteins in proportion to the other species. However, if 1-tm proteins are not considered, *P. vivax* and *P. yoelii* have more or less an equal percentage of transmembrane proteins. The percentage of proteins containing one or more transmembrane helices, is comparable to that reported for other species (Krogh *et al.*, 2001). It should be kept in mind that a weakness of transmembrane predictors is that proteins with signal peptides are often falsely predicted to be a transmembrane protein. Many of the false positively predicted membrane proteins would be resembled in the portion of predicted 1-tm proteins. Therefore, a reasonable assumption is that many of the 1-tm proteins actually contain a signal peptide which is mistaken for a transmembrane helix.

Coiled-coils content in P. falciparum and comparison with eukaryotes

P. falciparum and *P. vivax* have approximately the same percentage of coiled-coil proteins, while *P. yoelii* contains a smaller percentage of coiled-coil proteins. The amounts

of coiled-coil proteins are comparable in all three species. Therefore it seems that *P. yoelii* might have a large amount of incorrectly predicted proteins. The proportions of coiled-coil proteins in *P. falciparum* and *P. vivax* (approximately 10%) are comparable to the percentages previously predicted for eukaryotes.

Database coverage for P. falciparum, P. vivax and P. yoelii

The relative percentages of proteins covered by the PDB and by Pfam are the most for *P. falciparum* and the least for *P. yoelii*. *P. yoelii* had approximately the same amount of matches in these databases than *P. falciparum,* but has more predicted proteins. *P. yoelii* and *P. vivax* had more PRINTS and PROSITE hits, but a closer investigation revealed that the higher amount of matches are due to patterns and fingerprints with a high probability of occurrence which are repeatedly matched. *P. falciparum* proteins are better characterized since it was the first genome to be sequenced of the three species. Moreover, *P. falciparum* is responsible for the most virulent form of malaria and is therefore a higher priority than *P. yoelii* and *P. vivax*. *P. yoelli* had slightly more matches in Pfam than *P. falciparum*. This fact reinforces the notion that *P. yoelii* proteins have less inserts and therefore are more similar to characterized proteins families. Therefore, *P. yoelii* orthologs can possibly be used to correctly identify the protein family of *P. falciparum* proteins when the *P. falciparum* proteins do not show enough similarity to the family due to inserts.

3.4.3. *P. falciparum* targets for structural studies

Experimental structure determination

Proteins in *P. falciparum* most suitable for experimental structure determination have been putatively identified and prioritized. These include 178 proteins with little disorder, transmembrane, coiled-coils, signal peptide and low complexity content. In addition, it is predicted that approximately 1 200 *P. falciparum* protein structures would not be solved

by conventional experimental procedures due to predicted disorder and transmembrane content covering more than 80% of the sequence length. More than 80% of all the proteins identified, are less than 500 amino acids long. The average length of putatively identified targets is 354 amino acids whereas the average length for proteins with determined crystal structures is 424. The average predicted disorder and transmembrane content for proteins with determined crystal structures is 28.4%. Therefore, the disorder/transmembrane helix-content cut-off limit of 30% for identifying targets, is reasonable. In addition, the average lengths, IP, EC, improbability of expression in inclusion bodies and the total unwanted content are similar for the two groups of proteins, further supporting the identification method used.

Targets for homology modeling

Proteins suitable for homology modeling were identified based on similar proteins in the PDB. Targets were classified into groups according to the percent identity, e-value and coverage of the PDB match. 373 proteins were identified for which high quality models can possibly be built. Therefore, these proteins consist of a single domain. Another 41 proteins were identified for which good homology models can be built. The average length of these proteins is 369. Therefore, most of these proteins are single-domain proteins. Although there are more proteins which are suitable for homology modeling, these proteins are not included in the target list since domains should first be identified for these proteins. These domains should then fulfill the same criteria as was set here for the whole sequences.

Putative targets for in silico docking studies

A selection of 244 proteins were made which have small molecule binding sites and which have either crystal structures or are very similar to a protein with a determined crystal structure. These proteins are therefore suitable for *in silico* docking studies. Investigating these proteins through the web interface, more information regarding the

types of molecules binding to these sites, can be gained by following links to PDB, CDD, PubChem and Superligands. The structures of these proteins should also be compared to their human homologs. If there are structural differences between the homologs, putative drug targets can be identified.

3.5. Conclusions

In this chapter, different structural feature contents were compared between *P. yoelii*, *P. vivax* and *P. falciparum* predicted proteins. *P. falciparum* shows more variation in protein length and has longer proteins on average. *P. falciparum* also contains more predicted low complexity and disorder content than the other two species. Therefore, *P. falciparum* proteins probably contain more inserts than the other two species. Comparisons of *P. falciparum* proteins to *P. yoelii* and *P. vivax* proteins can be used to putatively identify inserts. The three species contain approximately the same amounts of different transmembrane proteins, excluding 1-tm and 2-tm proteins. The three species have similar amounts of coiled-coil and disorder-containing proteins, and similar amounts of proteins represented in family and structure databases. However, the relative percentages of proteins in the proteomes differ considerably for *P. yoelii* as a result of the large amount of total predicted proteins compared to *P. falciparum* and *P. vivax*. Therefore, *P. yoelii* probably has a portion of incorrectly predicted proteins. Interacting proteins in *P. falciparum*, contain a higher predicted disorder content than non-interacting proteins. Furthermore, most proteins interacting with ten or more proteins have a disorder content in the range of 60-100%, whereas the disorder contents of proteins interacting with one protein, are evenly distributed between 1-100%.

Experimental structure determination is difficult and time consuming, therefore it is useful to identify proteins which will contribute most to exploring the structure space in the *P. falciparum* proteome. Such proteins can be identified by finding the proteins least represented in the PDB. It is also logical to identify those proteins which are likely to be less problematic during the various steps of the experimental procedures. Complicating

features include large transmembrane domains and a high percentage of the sequence forming either disordered, coiled-coil or low complexity regions. Proteins suitable for experimental structure determination, homology modeling and *in silico* docking studies were putatively identified. Functionally characterized targets were separated from uncharacterized proteins. The identified targets were compared to proteins with solved crystal structures and the average predicted disorder and transmembrane helix content are similar for the two groups of proteins. Furthermore, the two groups of proteins have similar average iso-electric points, extinction coefficients and improbabilities of expression in inclusion bodies. These lists will provide the malaria research community with high priority targets suitable for experimental structure determination. Determination of the structures of these proteins will facilitate the identification of a wide range of novel drug targets. A group of 1 200 proteins were identified which are either wholly disordered or consist completely of a large transmembrane domain. Therefore, the structures of these proteins will possibly never be solved by X-ray crystallization. Many proteins contain domains whose structures can be solved together with largely disordered domains. Therefore, domain identification is very important for the determination of target domains. Domain identification should be incorporated into target identification of suitable proteins for structural studies.

Chapter 4

Concluding discussion

Malaria is responsible for millions of deaths per year. *Plasmodium* species continue to develop resistance against existing anti-malarial drugs, therefore new drugs need to be developed. An important part of the drug development process is the identification of novel drug targets. Once novel drug targets have been identified, inhibitors can be designed from the protein structure. However, knowledge of protein function and 3D structure is necessary for the identification of therapeutic targets. Experimental structure determination of *P. falciparum* proteins is especially difficult as a result of the abundance of low complexity, inserts and unstructured regions. Although some proteins can be modeled through homology, most malarial proteins do not have sufficient similarity to proteins in the PDB. In response to the lack of structural characterization in *P. falciparum*, integrated structural feature annotation was performed on the proteome. These annotations were used to putatively identify proteins which are suitable for various structural studies.

Structural annotations for *P. falciparum*, *P. vivax* and *P. yoelii* were performed on a Linux cluster and the results were stored in a PostgreSQL database. All the proteins were searched for similarities to proteins or protein families in the PDB, Pfam, PRINTS, PROSITE and Superfamily databases. Secondary structure, transmembrane helices, disorder, low complexity, coiled-coils and small molecule interaction predictions were made for the proteins. In addition, selected proteins were subjected to threading. Protein-protein interactions and proteins exported to the RBC were annotated from literature. These annotations provide a comprehensive overview of the features predicted to

be present in every protein and can be viewed through a web interface which graphically illustrates the locations of features within the sequence. The structural annotations are useful for designing experiments for function and structure determination. PLASMODB contains integrated information about the genes and predicted proteins from all the *Plasmodium* species. Several of the annotations included in the system presented here, are also included in PLASMODB. However, the focus of this annotation was on structural features necessary to identify suitable proteins for structural studies. Threading and predictions of disorder, coiled-coil and small molecule interactions are not included in PLASMODB. In order to select groups of proteins which fulfill certain criteria with regard to structural and functional features, a query tool was constructed to automatically build PostgreSQL queries. By using this tool, criteria can be set which include the presence or absence of all the predicted features. The sequence identity to and e-values of PDB matches can be specified. Additional criteria with regard to protein length and the availability of a crystal structure can be set.

Validation studies revealed that Threader results are reliable when Z-scores of at least 3.95 are obtained. For the two proteins investigated, secondary structure, disorder and transmembrane predictions correlated with experimental evidence. The features which least overlapped with aligned PDB matches, were transmembrane helix and disorder predictions. Therefore, it was concluded that these predictions are the most useful for the identification of suitable candidates for experimental structure determination. Furthermore, scanning against an integrated family database such as InterPro, is more efficient and provides better coverage of protein domain families.

Analysis of the results obtained for the three species, revealed that *P. falciparum* proteins tend to be longer and vary more in length than the other two species. *P. falciparum* proteins also contain more predicted low complexity and disorder content than proteins from *P. yoelii* and *P. vivax*. Therefore, it was concluded that *P. falciparum* proteins contain more inserts. Additionally, comparisons to orthologous proteins from *P. vivax* and

P. yoelii might reveal insert locations. Approximately equal amounts of proteins from *P. falciparum*, *P. vivax* and *P. yoelii* contain coiled-coils, disorder and 3-tm to 15-tm helices. In addition, equal amounts of proteins had matches in the different databases excluding high probability occurring matches. However, the relative percentages of proteins in the proteomes differ significantly for *P. yoelii* as a result of the large number of total predicted proteins compared to *P. falciparum* and *P. vivax*. Therefore, *P. yoelii* probably has a portion of incorrectly predicted proteins. Protein interacting proteins contain a higher percentage of predicted disordered residues than non-interacting proteins. Proteins interacting with 10 or more other proteins have a disordered content concentrated in the range of 60-100%, while the disorder distribution for proteins having only one interacting partner, was more evenly spread.

In order to increase the efficiency of structural genomics for malaria species, suitable targets for experimental structure determination were putatively identified. Targets likely to have novel structures were selected in order to identify those proteins which will contribute most to the exploration of the structural space in *P. falciparum*. Therefore, proteins which are least represented in the PDB, based on the BLAST-PDB search, were selected. Subsequently, proteins were eliminated based on predicted transmembrane helix, disorder, low complexity, coiled-coils and signal peptide content. Finally, functionally characterized proteins and protein-interacting proteins were separated from uncharacterized targets. Comparisons of the 178 identified targets for experimental structure determination to proteins with crystal structures, revealed that extinction coefficients, improbabilities of expression in inclusion bodies and iso-electric points were similar for the two groups of proteins. In addition, the average predicted disorder and transmembrane helix contents were similar, but slightly lower for the predicted targets. Therefore, the cut-off limits set for unfavourable structural features for crystallization were sufficient. Targets for homology modeling were principally based on similarity to proteins in the PDB. Thresholds were set for the e-values of matches, the percent sequence identity of the match and the query sequence and the percentage of the query sequence which

was covered by the match. A group of 373 proteins had matches which fulfilled the criteria. These are probably proteins with a single domain since a high percentage of query sequence needed to be covered by the PDB match. Domains from multi-domain proteins suitable for homology modeling can only be identified once domain boundaries have been predicted for all the proteins. The number of proteins which consist more or less completely out of either disordered regions or transmembrane regions is 1 200. The structures of these proteins could probably not be determined by X-ray crystallization. Finally, 197 targets for *in silico* docking were identified based on predicted small molecule interactions and the availability of a 3D structure.

Conclusively, the structural annotation presented here provides malaria researchers with a tool to select specific groups of proteins based on predicted structural and functional features. It provides lists of putative targets for various structural studies. In future, it might also facilitate the selection of putative drug targets by including additional analysis such as similarity searches against human proteins. The structural annotation can be extended to other species.

Appendix

Psipred and VSL2 output files for PFE0660c

```
# PSIPRED HFORMAT (PSIPRED V2.5 by David Jones)
Conf:  988512468883563887998599789999999716904564028745999999997999
Pred:  CCCCEECCCCCHHHCCCEEEECCCHHHHHHHHHHHHCCEEEEECCCCEEEEEEEECCEEEE
Am-A : MDNLLRHLKISKEQITPVVLVVGDPGRVDKIKVVCDSYVDLAYNREYKSVECHYKGQKFL
---------------10--------20--------30--------40--------50--------60
Conf:  984688887999999999843998799950205537355546748997210104684333
Pred:  EEECCCCHHHHHHHHHHHHHCCCCEEEEEECCCCCCCCCCCCCCCEEEEEHHCCCCCCCCC
Am-A : CVSHGVGSAGCAVCFEELCQNGAKVIIRAGSCGSLQPDLIKRGDICICNAAVREDRVSHL
---------------70--------80--------90-------100-------110-------120
Conf:  258656789998999999999997699489999987686357721589999998649939
Pred:  CCCCCCCCCCCHHHHHHHHHHHHHCCCCEEEEEEEECCCCCCCCCHHHHHHHHHHCCCCE
Am-A : LIHGDFPAVGDFDVYDTLNKCAQELNVPVFNGISVSSDMYYPNKIIPSRLEDYSKANAAV
-------------130-------140-------150-------160-------170-------180
Conf:  97267899999998089889999851777664313420689999999999999999998
Pred:  EECCHHHHHHHHHHHCCCEEEEEEEECCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHH
Am-A : VEMELATLMVIGTLRKVKTGGILIVDGCPFKWDEGDFDNNLVPHQLENMIKIALGACAKL
-------------190-------200--------210-------220-------230-------240
Conf:  74149
Pred:  HHHCC
Am-A : ATKYA

VSL2 Predictor of Intrinsically Disordered Regions
Center for Information Science and Technology
Temple University, Philadelphia, PA
Predicted Disordered Regions:
1-7
```

212-221

240-245

Prediction Scores:

==========

NO. RES. PREDICTION DISORDER

----

1 M 0.877276 D

2 D 0.853825 D

3 N 0.848926 D

4 L 0.816596 D

5 L 0.751564 D

6 R 0.643868 D

7 H 0.551074 D

212 W 0.593946 D

213 D 0.673446 D

214 E 0.733506 D

215 G 0.765616 D

216 D 0.761963 D

217 F 0.746079 D

218 D 0.729025 D

219 N 0.698407 D

220 N 0.627326 D

221 L 0.530775 D

240 L 0.511101 D

241 A 0.645568 D

242 T 0.770339 D

243 K 0.833105 D

244 Y 0.878222 D

245 A 0.917207 D

==========

Bibliography

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) Basic local alignment search tool. *J Mol Biol* **215**, 3, 403–410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and LipmanG, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 17, 3389–3402.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 1, 25–29.

Attwood, T. K., Bradley, P., Flower, D. R., Gaulton, A., Maudling, N., Mitchell, A. L., Moulton, G., Nordle, A., Paine, K., Taylor, P., Uddin, A. and Zygouri, C. (2003) PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res* **31**, 1, 400–402.

Bader, S., Kortholt, A., Snippe, H. and Haastert, P. J. M. V. (2006) DdPDE4, a novel cAMP-specific phosphodiesterase at the surface of *dictyostelium* cells. *J Biol Chem* **281**, 29, 20018–20026.

Bairoch, A. (1991) PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Res* **19 Suppl**, 2241–2245.

Bairoch, A., Boeckmann, B., Ferro, S. and Gasteiger, E. (2004) Swiss-Prot: juggling between evolution and stability. *Brief Bioinform* **5**, 1, 39–55.

Bateman, A. and Haft, D. H. (2002) HMM-based databases in InterPro. *Brief Bioinform* **3**, 3, 236–245.

Bendtsen, J. D., Nielsen, H., von Heijne, G. and Brunak, S. (2004) Improved prediction

of signal peptides: SignalP 3.0. *J Mol Biol* **340**, 4, 783–795.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res* **28**, 1, 235–242.

Blake, J. A., Eppig, J. T., Bult, C. J., Kadin, J. A., Richardson, J. E. and Group, M. G. D. (2006) The Mouse Genome Database (MGD): updates and enhancements. *Nucleic Acids Res* **34**, Database issue, D562–D567.

Boscott, P. E., Barton, G. J. and Richards, W. G. (1993) Secondary structure prediction for modelling by homology. *Protein Eng* **6**, 3, 261–266.

Brooks, D. R., Wang, P., Read, M., Watkins, W. M., Sims, P. F. and Hyde, J. E. (1994) Sequence variation of the hydroxymethyldihydropterin pyrophosphokinase: dihydropteroate synthase gene in lines of the human malaria parasite, *Plasmodium falciparum*, with differing resistance to sulfadoxine. *Eur J Biochem* **224**, 2, 397–405.

Carlton, J. (2003) The *Plasmodium vivax* genome sequencing project. *Trends Parasitol* **19**, 5, 227–231.

Carlton, J. M., Angiuoli, S. V., Suh, B. B., Kooij, T. W., Pertea, M., Silva, J. C., Ermolaeva, M. D., Allen, J. E., Selengut, J. D., Koo, H. L., Peterson, J. D., Pop, M., Kosack, D. S., Shumway, M. F., Bidwell, S. L., Shallom, S. J., van Aken, S. E., Riedmuller, S. B., Feldblyum, T. V., Cho, J. K., Quackenbush, J., Sedegah, M., Shoaibi, A., Cummings, L. M., Florens, L., Yates, J. R., Raine, J. D., Sinden, R. E., Harris, M. A., Cunningham, D. A., Preiser, P. R., Bergman, L. W., Vaidya, A. B., van Lin, L. H., Janse, C. J., Waters, A. P., Smith, H. O., White, O. R., Salzberg, S. L., Venter, J. C., Fraser, C. M., Hoffman, S. L., Gardner, M. J. and Carucci, D. J. (2002) Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. *Nature* **419**, 6906, 512–519.

Carter, P., Liu, J. and Rost, B. (2003) PEP: Predictions for Entire Proteomes. *Nucleic Acids Res* **31**, 1, 410–413.

Chen, L., Oughtred, R., Berman, H. M. and Westbrook, J. (2004) TargetDB: a target registration database for structural genomics projects. *Bioinformatics* **20**, 16, 2860–2862.

Chigira, S., Sugita, K., Kita, K., Sugaya, S., Arase, Y., Ichinose, M., Shirasawa, H. and Suzuki, N. (2003) Increased expression of the Huntingtin interacting protein-1 gene in cells from Hutchinson Gilford Syndrome (Progeria) patients and aged donors. *J Gerontol A Biol Sci Med Sci* **58**, 10, B873–B878.

Chothia, C. (1992) Proteins. One thousand families for the molecular biologist. *Nature* **357**, 6379, 543–544.

Chothia, C. and Lesk, A. M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J* **5**, 4, 823–826.

Cuthbertson, J. M., Doyle, D. A. and Sansom, M. S. P. (2005) Transmembrane helix prediction: a comparative evaluation and analysis. *Protein Eng Des Sel* **18**, 6, 295–308.

D'Angelo, M. A., Sanguineti, S., Reece, J. M., Birnbaumer, L., Torres, H. N. and Flawia, M. M. (2004) Identification, characterization and subcellular localization of TcPDE1, a novel cAMP-specific phosphodiesterase from *Trypanosoma cruzi*. *Biochem J* **378**, Pt 1, 63–72.

Date, S. V. and Stoeckert, C. J. (2006) Computational modeling of the *Plasmodium falciparum* interactome reveals protein function on a genome-wide scale. *Genome Res* **16**, 4, 542–549.

Dayhoff, M. O., Barker, W. C. and McLaughlin, P. J. (1974) Inferences from protein and nucleic acid sequences: early molecular evolution, divergence of kingdoms and rates of change. *Orig Life* **5**, 3, 311–330.

de Beer, T. A. P., Louw, A. I. and Joubert, F. (2006) Elucidation of sulfadoxine resistance with structural models of the bifunctional *Plasmodium falciparum* dihydropterin pyrophosphokinase-dihydropteroate synthase. *Bioorg Med Chem* **14**, 13, 4433–4443.

Dosztanyi, Z., Csizmok, V., Tompa, P. and Simon, I. (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21**, 16, 3433–3434.

Dunker, A. K., Obradovic, Z., Romero, P., Garner, E. C. and Brown, C. J. (2000) Intrinsic protein disorder in complete genomes. *Genome Inform Ser Workshop Genome Inform* **11**, 161–171.

Edgar, R. C. and Sjolander, K. (2004) A comparison of scoring functions for protein sequence profile alignment. *Bioinformatics* **20**, 8, 1301–1308.

Eisenberg, D., Marcotte, E. M., Xenarios, I. and Yeates, T. O. (2000) Protein function in the post-genomic era. *Nature* **405**, 6788, 823–826.

Finn, R. D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S. R., Sonnhammer, E. L. L. and Bateman, A. (2006) Pfam: clans, web tools and services. *Nucleic Acids Res* **34**, Database issue, D247–D251.

Frishman, D. (2002) Knowledge-based selection of targets for structural genomics. *Protein Eng* **15**, 3, 169–183.

Frishman, D., Mokrejs, M., Kosykh, D., Kastenmuller, G., Kolesov, G., Zubrzycki, I., Gruber, C., Geier, B., Kaps, A., Albermann, K., Volz, A., Wagner, C., Fellenberg, M., Heumann, K. and Mewes, H.-W. (2003) The PEDANT genome database. *Nucleic Acids Res* **31**, 1, 207–211.

Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., Carlton, J. M., Pain, A., Nelson, K. E., Bowman, S., Paulsen, I. T., James, K., Eisen, J. A., Rutherford, K., Salzberg, S. L., Craig, A., Kyes, S., Chan, M.-S., Nene, V., Shallom, S. J., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M. W., Vaidya, A. B., Martin, D. M. A., Fairlamb, A. H., Fraunholz, M. J., Roos, D. S., Ralph, S. A., McFadden, G. I., Cummings, L. M., Subramanian, G. M., Mungall, C., Venter, J. C., Carucci, D. J., Hoffman, S. L., Newbold, C., Davis, R. W., Fraser, C. M. and Barrell, B. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 6906, 498–511.

Gelbart, W. M., Crosby, M., Matthews, B., Rindone, W. P., Chillemi, J., Twombly, S. R., Emmert, D., Ashburner, M., Drysdale, R. A., Whitfield, E., Millburn, G. H., de Grey, A., Kaufman, T., Matthews, K., Gilbert, D., Strelets, V. and Tolstoshev, C. (1997) FlyBase: a *Drosophila* database. The FlyBase consortium. *Nucleic Acids Res* **25**, 1, 63–66.

Ginalski, K., Grishin, N. V., Godzik, A. and Rychlewski, L. (2005) Practical lessons from

protein structure prediction. *Nucleic Acids Res* **33**, 6, 1874–1891.

Ginalski, K. and Rychlewski, L. (2003) Detection of reliable and unexpected protein fold predictions using 3D-Jury. *Nucleic Acids Res* **31**, 13, 3291–3292.

Gough, J., Karplus, K., Hughey, R. and Chothia, C. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J Mol Biol* **313**, 4, 903–919.

Gouzy, J., Corpet, F. and Kahn, D. (1999) Whole genome protein domain analysis using a new method for domain clustering. *Comput Chem* **23**, 3-4, 333–340.

Gowri, V. S., Pandit, S. B., Karthik, P. S., Srinivasan, N. and Balaji, S. (2003) Integration of related sequences with protein three-dimensional structural families in an updated version of PALI database. *Nucleic Acids Res* **31**, 1, 486–488.

Grana, O., Eyrich, V. A., Pazos, F., Rost, B. and Valencia, A. (2005) EVAcon: a protein contact prediction evaluation service. *Nucleic Acids Res* **33**, Web Server issue, W347–W351.

Haft, D. H., Selengut, J. D. and White, O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res* **31**, 1, 371–373.

Haldar, K., Mohandas, N., Samuel, B. U., Harrison, T., Hiller, N. L., Akompong, T. and Cheresh, P. (2002) Protein and lipid trafficking induced in erythrocytes infected by malaria parasites. *Cell Microbiol* **4**, 7, 383–395.

Haynes, C., Oldfield, C. J., Ji, F., Klitgord, N., Cusick, M. E., Radivojac, P., Uversky, V. N., Vidal, M. and Iakoucheva, L. M. (2006) Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS Comput Biol* **2**, 8, e100.

Henikoff, J. G., Greene, E. A., Pietrokovski, S. and Henikoff, S. (2000) Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res* **28**, 1, 228–230.

Henikoff, S. and Henikoff, J. G. (1991) Automated assembly of protein blocks for database searching. *Nucleic Acids Res* **19**, 23, 6565–6572.

Herrera, F. E., Zucchelli, S., Jezierska, A., Lavina, Z. S., Gustincich, S. and Carloni, P. (2007) On the oligomeric state of DJ-1 protein and its mutants associated with Parkinson's Disease: a combined computational and in vitro study. *J Biol Chem.*

Hiller, N. L., Bhattacharjee, S., van Ooij, C., Liolios, K., Harrison, T., Lopez-Estrano, C. and Haldar, K. (2004) A host-targeting signal in virulence proteins reveals a secretome in malarial infection. *Science* **306**, 5703, 1934–1937.

Iakoucheva, L. M. and Dunker, A. K. (2003) Order, disorder, and flexibility: prediction from protein sequence. *Structure* **11**, 11, 1316–1317.

Jennings, A. J., Edge, C. M. and Sternberg, M. J. (2001) An approach to improving multiple alignments of protein sequences using predicted secondary structure. *Protein Eng* **14**, 4, 227–231.

Jensen, L. J., Gupta, R., Staerfeldt, H.-H. and Brunak, S. (2003) Prediction of human protein function according to Gene Ontology categories. *Bioinformatics* **19**, 5, 635–642.

Jones, D. and Hadley, J. C. (2000) *Threading methods for protein structure prediction. In Bioinformatics, sequence, structure and databanks* Oxford University Press, Oxford, UK.

Jones, D. T. (1999*a*) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J Mol Biol* **287**, 4, 797–815.

Jones, D. T. (1999*b*) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* **292**, 2, 195–202.

Jones, D. T., Miller, R. T. and Thornton, J. M. (1995) Successful protein fold recognition by optimal sequence threading validated by rigorous blind testing. *Proteins* **23**, 3, 387–397.

Jones, D. T., Taylor, W. R. and Thornton, J. M. (1992) A new approach to protein fold recognition. *Nature* **358**, 6381, 86–89.

Jones, D. T., Tress, M., Bryson, K. and Hadley, C. (1999) Successful recognition of protein folds using threading methods biased by sequence similarity and predicted secondary structure. *Proteins* **Suppl 3**, 104–111.

Juncker, A. S., Willenbrock, H., Heijne, G. V., Brunak, S., Nielsen, H. and Krogh, A. (2003) Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci* **12**, 8, 1652–1662.

Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pat-

tern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 12, 2577–2637.

Kahsay, R. Y., Gao, G. and Liao, L. (2005) An improved hidden Markov model for transmembrane protein detection and topology prediction and its applications to complete genomes. *Bioinformatics* **21**, 9, 1853–1858.

Kall, L., Krogh, A. and Sonnhammer, E. L. L. (2004) A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* **338**, 5, 1027–1036.

Karplus, K., Barrett, C. and Hughey, R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **14**, 10, 846–856.

Krogh, A., Brown, M., Mian, I. S., Sjolander, K. and Haussler, D. (1994) Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol* **235**, 5, 1501–1531.

Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**, 3, 567–580.

LaCount, D. J., Vignali, M., Chettier, R., Phansalkar, A., Bell, R., Hesselberth, J. R., Schoenfeld, L. W., Ota, I., Sahasrabudhe, S., Kurschner, C., Fields, S. and Hughes, R. E. (2005) A protein interaction network of the malaria parasite *Plasmodium falciparum*. *Nature* **438**, 7064, 103–107.

Lemer, C. M., Rooman, M. J. and Wodak, S. J. (1995) Protein structure prediction by threading methods: evaluation of current techniques. *Proteins* **23**, 3, 337–355.

Letunic, I., Copley, R. R., Pils, B., Pinkert, S., Schultz, J. and Bork, P. (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res* **34**, Database issue, D257–D260.

Lin, H.-N., Chang, J.-M., Wu, K.-P., Sung, T.-Y. and Hsu, W.-L. (2005) HYPROSP II–a knowledge-based hybrid method for protein secondary structure prediction based on local prediction confidence. *Bioinformatics* **21**, 15, 3227–3233.

Liu, J., Hegyi, H., Acton, T. B., Montelione, G. T. and Rost, B. (2004) Automatic target selection for structural genomics on eukaryotes. *Proteins* **56**, 2, 188–200.

Liu, J. and Rost, B. (2001) Comparing function and structure between entire proteomes. *Protein Sci* **10**, 10, 1970–1979.

Lupas, A., Dyke, M. V. and Stock, J. (1991) Predicting coiled coils from protein sequences. *Science* **252**, 5010, 1162–1164.

Madera, M. and Gough, J. (2002) A comparison of profile hidden Markov model procedures for remote homology detection. *Nucleic Acids Res* **30**, 19, 4321–4328.

Madera, M., Vogel, C., Kummerfeld, S. K., Chothia, C. and Gough, J. (2004) The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res* **32**, Database issue, D235–D239.

Marchler-Bauer, A., Anderson, J. B., Cherukuri, P. F., DeWeese-Scott, C., Geer, L. Y., Gwadz, M., He, S., Hurwitz, D. I., Jackson, J. D., Ke, Z., Lanczycki, C. J., Liebert, C. A., Liu, C., Lu, F., Marchler, G. H., Mullokandov, M., Shoemaker, B. A., Simonyan, V., Song, J. S., Thiessen, P. A., Yamashita, R. A., Yin, J. J., Zhang, D. and Bryant, S. H. (2005) CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Res* **33**, Database issue, D192–D196.

Marsden, R. L., McGuffin, L. J. and Jones, D. T. (2002) Rapid protein domain assignment from amino acid sequence using predicted secondary structure. *Protein Sci* **11**, 12, 2814–2824.

Marti, M., Good, R. T., Rug, M., Knuepfer, E. and Cowman, A. F. (2004) Targeting malaria virulence and remodeling proteins to the host erythrocyte. *Science* **306**, 5703, 1930–1933.

McClatchey, A. I. (2003) Merlin and ERM proteins: unappreciated roles in cancer development? *Nat Rev Cancer* **3**, 11, 877–883.

McDonnell, A. V., Jiang, T., Keating, A. E. and Berger, B. (2006) Paircoil2: improved prediction of coiled coils from sequence. *Bioinformatics* **22**, 3, 356–358.

Meinel, T., Krause, A., Luz, H., Vingron, M. and Staub, E. (2005) The SYSTERS Protein Family Database in 2005. *Nucleic Acids Res* **33**, Database issue, D226–D229.

Michalsky, E., Dunkel, M., Goede, A. and Preissner, R. (2005) SuperLigands - a database of ligand structures derived from the Protein Data Bank. *BMC Bioinformatics* **6**, 122.

Miller, L. H., Baruch, D. I., Marsh, K. and Doumbo, O. K. (2002) The pathogenic basis of malaria. *Nature* **415**, 6872, 673–679.

Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bradley, P., Bork, P., Bucher, P., Cerutti, L., Copley, R., Courcelle, E., Das, U., Durbin, R., Fleischmann, W., Gough, J., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lonsdale, D., Lopez, R., Letunic, I., Madera, M., Maslen, J., McDowall, J., Mitchell, A., Nikolskaya, A. N., Orchard, S., Pagni, M., Ponting, C. P., Quevillon, E., Selengut, J., Sigrist, C. J. A., Silventoinen, V., Studholme, D. J., Vaughan, R. and Wu, C. H. (2005) InterPro, progress and status in 2005. *Nucleic Acids Res* **33**, Database issue, D201–D205.

Murzin, A. G., Brenner, S. E., Hubbard, T. and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* **247**, 4, 536–540.

Pandit, S. B., Gosar, D., Abhiman, S., Sujatha, S., Dixit, S. S., Mhatre, N. S., Sowdhamini, R. and Srinivasan, N. (2002) SUPFAM–a database of potential protein superfamily relationships derived by comparing sequence-based and structure-based families: implications for structural genomics and function annotation in genomes. *Nucleic Acids Res* **30**, 1, 289–293.

Pearl, F., Todd, A., Sillitoe, I., Dibley, M., Redfern, O., Lewis, T., Bennett, C., Marsden, R., Grant, A., Lee, D., Akpor, A., Maibaum, M., Harrison, A., Dallman, T., Reeves, G., Diboun, I., Addou, S., Lise, S., Johnston, C., Sillero, A., Thornton, J. and Orengo, C. (2005) The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Res* **33**, Database issue, D247–D251.

Peng, K., Radivojac, P., Vucetic, S., Dunker, A. K. and Obradovic, Z. (2006) Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics* **7**, 1, 208.

Peterson, D. S., Walliker, D. and Wellems, T. E. (1988) Evidence that a point mutation in dihydrofolate reductase-thymidylate synthase confers resistance to pyrimethamine

in *falciparum* malaria. *Proc Natl Acad Sci U S A* **85**, 23, 9114–9118.

Puls, I., Jonnakuty, C., LaMonte, B. H., Holzbaur, E. L. F., Tokito, M., Mann, E., Floeter, M. K., Bidus, K., Drayna, D., Oh, S. J., Brown, R. H., Ludlow, C. L. and Fischbeck, K. H. (2003) Mutant dynactin in motor neuron disease. *Nat Genet* **33**, 4, 455–456.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R. and Lopez, R. (2005) InterProScan: protein domains identifier. *Nucleic Acids Res* **33**, Web Server issue, W116–W120.

Rayment, I., Holden, H. M., Whittaker, M., Yohn, C. B., Lorenz, M., Holmes, K. C. and Milligan, R. A. (1993) Structure of the actin-myosin complex and its implications for muscle contraction. *Science* **261**, 5117, 58–65.

Riley, M. L., Schmidt, T., Artamonova, I. I., Wagner, C., Volz, A., Heumann, K., Mewes, H.-W. and Frishman, D. (2007) PEDANT genome database: 10 years online. *Nucleic Acids Res* **35**, Database issue, D354–D357.

Romero, P., Obradovic, Z. and Dunker, A. K. (2004) Natively disordered proteins: functions and predictions. *Appl Bioinformatics* **3**, 2-3, 105–113.

Romero, P., Obradovic, Z., Kissinger, C. R., Villafranca, J. E., Garner, E., Guilliot, S. and Dunker, A. K. (1998) Thousands of proteins likely to have long disordered regions. *Pac Symp Biocomput* 437–448.

Rose, A., Schraegle, S. J., Stahlberg, E. A. and Meier, I. (2005) Coiled-coil protein composition of 22 proteomes–differences and common themes in subcellular infrastructure and traffic control. *BMC Evol Biol* **5**, 66.

Rost, B., Casadio, R., Fariselli, P. and Sander, C. (1995) Transmembrane helices predicted at 95521–533.

Rozmajzl, P. J., Kimura, M., Woodrow, C. J., Krishna, S. and Meade, J. C. (2001) Characterization of P-type ATPase 3 in *Plasmodium falciparum*. *Mol Biochem Parasitol* **116**, 2, 117–126.

Sarma, G. N., Savvides, S. N., Becker, K., Schirmer, M., Schirmer, R. H. and Karplus, P. A. (2003) Glutathione reductase of the malarial parasite *Plasmodium falciparum*:

crystal structure and inhibitor development. *J Mol Biol* **328**, 4, 893–907.

Schaffer, A. A., Wolf, Y. I., Ponting, C. P., Koonin, E. V., Aravind, L. and Altschul, S. F. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics* **15**, 12, 1000–1011.

Schnick, C., Robien, M. A., Brzozowski, A. M., Dodson, E. J., Murshudov, G. N., Anderson, L., Luft, J. R., Mehlin, C., Hol, W. G. J., Brannigan, J. A. and Wilkinson, A. J. (2005) Structures of *Plasmodium falciparum* purine nucleoside phosphorylase complexed with sulfate and its natural substrate inosine. *Acta Crystallogr D Biol Crystallogr* **61**, Pt 9, 1245–1254.

Servant, F., Bru, C., Carrere, S., Courcelle, E., Gouzy, J., Peyruc, D. and Kahn, D. (2002) ProDom: automated clustering of homologous domains. *Brief Bioinform* **3**, 3, 246–251.

Shakur, Y., Takeda, K., Kenan, Y., Yu, Z. X., Rena, G., Brandt, D., Houslay, M. D., Degerman, E., Ferrans, V. J. and Manganiello, V. C. (2000) Membrane localization of cyclic nucleotide phosphodiesterase 3 (PDE3). Two N-terminal domains are required for the efficient targeting to, and association of, PDE3 with endoplasmic reticulum. *J Biol Chem* **275**, 49, 38749–38761.

Shortle, D. (1997) Structure prediction: folding proteins by pattern recognition. *Curr Biol* **7**, 3, R151–R154.

Stoeckert, C. J., Fischer, S., Kissinger, J. C., Heiges, M., Aurrecoechea, C., Gajria, B. and Roos, D. S. (2006) PlasmoDB v5: new looks, new genomes. *Trends Parasitol* **22**, 12, 543–546.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J. and Natale, D. A. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41.

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R.,

Diemer, K., Muruganujan, A. and Narechania, A. (2003) PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res* **13**, 9, 2129–2141.

Tompa, P. (2005) The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett* **579**, 15, 3346–3354.

Tusnady, G. E., Dosztanyi, Z. and Simon, I. (2004) Transmembrane proteins in the Protein Data Bank: identification and classification. *Bioinformatics* **20**, 17, 2964–2972.

Velanker, S. S., Ray, S. S., Gokhale, R. S., Suma, S., Balaram, H., Balaram, P. and Murthy, M. R. (1997) Triosephosphate isomerase from *Plasmodium falciparum*: the crystal structure provides insights into antimalarial drug design. *Structure* **5**, 6, 751–761.

von Heijne, G. (1986) A new method for predicting signal sequence cleavage sites. *Nucleic Acids Res* **14**, 11, 4683–4690.

von Heijne, G. (1989) Control of topology and mode of assembly of a polytopic membrane protein by positively charged residues. *Nature* **341**, 6241, 456–458.

Vucetic, S., Brown, C. J., Dunker, A. K. and Obradovic, Z. (2003) Flavors of protein disorder. *Proteins* **52**, 4, 573–584.

Wallin, E. and von Heijne, G. (1998) Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci* **7**, 4, 1029–1038.

Ward, J. J., McGuffin, L. J., Bryson, K., Buxton, B. F. and Jones, D. T. (2004*a*) The DISOPRED server for the prediction of protein disorder. *Bioinformatics* **20**, 13, 2138–2139.

Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F. and Jones, D. T. (2004*b*) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* **337**, 3, 635–645.

Wickham, M. E., Rug, M., Ralph, S. A., Klonis, N., McFadden, G. I., Tilley, L. and Cowman, A. F. (2001) Trafficking and assembly of the cytoadherence complex in *Plasmodium falciparum*-infected human erythrocytes. *EMBO J* **20**, 20, 5636–5649.

Wolf, E., Kim, P. S. and Berger, B. (1997) MultiCoil: a program for predicting two- and three-stranded coiled coils. *Protein Sci* **6**, 6, 1179–1189.

Wootton, J. C. (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput Chem* **18**, 3, 269–285.

Wu, C. H., Apweiler, R., Bairoch, A., Natale, D. A., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Mazumder, R., O'Donovan, C., Redaschi, N. and Suzek, B. (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res* **34**, Database issue, D187–D191.

Wu, C. H., Huang, H., Arminski, L., Castro-Alvear, J., Chen, Y., Hu, Z.-Z., Ledley, R. S., Lewis, K. C., Mewes, H.-W., Orcutt, B. C., Suzek, B. E., Tsugita, A., Vinayaka, C. R., Yeh, L.-S. L., Zhang, J. and Barker, W. C. (2002) The Protein Information Resource: an integrated public resource of functional annotation of proteins. *Nucleic Acids Res* **30**, 1, 35–37.

Wu, C. H., Huang, H., Nikolskaya, A., Hu, Z. and Barker, W. C. (2004*a*) The iProClass integrated database for protein functional analysis. *Comput Biol Chem* **28**, 1, 87–96.

Wu, C. H., Nikolskaya, A., Huang, H., Yeh, L.-S. L., Natale, D. A., Vinayaka, C. R., Hu, Z.-Z., Mazumder, R., Kumar, S., Kourtesis, P., Ledley, R. S., Suzek, B. E., Arminski, L., Chen, Y., Zhang, J., Cardenas, J. L., Chung, S., Castro-Alvear, J., Dinkov, G. and Barker, W. C. (2004*b*) PIRSF: family classification system at the Protein Information Resource. *Nucleic Acids Res* **32**, Database issue, D112–D114.

Yeats, C., Maibaum, M., Marsden, R., Dibley, M., Lee, D., Addou, S. and Orengo, C. A. (2006) Gene3D: modelling protein structure, function and evolution. *Nucleic Acids Res* **34**, Database issue, D281–D284.

Yuthavong, Y., Yuvaniyama, J., Chitnumsub, P., Vanichtanankul, J., Chusacultanachai, S., Tarnchompoo, B., Vilaivan, T. and Kamchonwongpaisan, S. (2005) Malarial *Plasmodium falciparum* dihydrofolate reductase-thymidylate synthase: structural basis for antifolate resistance and development of effective inhibitors. *Parasitology* **130**, Pt 3, 249–259.

Zhang, Y., Yin, Y., Chen, Y., Gao, G., Yu, P., Luo, J. and Jiang, Y. (2003) PCAS–a precomputed proteome annotation database resource. *BMC Genomics* **4**, 1, 42.