In silico prediction of host-pathogen protein-protein interactions in the malaria parasite, *Plasmodium falciparum*

by

CHRISTIAAN ODENDAAL

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

13 September 2010

Declaration

I, Christiaan Odendaal declare that the thesis/dissertation, which I hereby submit for the degree MSc. Bioinformatics at the University of Pretoria, is my own work and that it has not previously been submitted by me for a degree at this or any other tertiary institution.

Christiaan Odendaal

20 July 2010

Acknowledgements

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

- To God for His guidance and assistance through wonderful people and for Him giving me this opportunity to study my MSc degree in Bioinformatics
- To Prof. Fourie Joubert, for his exceptional insight and guidance during this project and for his support and encouragement
- To my wife, Hester, for her love and support during the completion of this project
- To all my colleagues in the Bioinformatics and Computational unit of the University of Pretoria, especially Tjaart, Claudia and Michal
- To the Department of Biochemistry and the Bioinformatics and Computational Unit of the University of Pretoria for providing their facilities and a sound academic environment
- To the National Bioinformatics Network (NBN) and the University of Pretoria for awarding me bursaries which enabled me to undertake an MSc degree

Summary

Malaria claims millions of lives annually. This global killer causes approximately 2.7 million annual deaths worldwide; addressing this problem has become more and more crucial. Due to pathogen evolution no efficient vaccine for treatment of malaria currently exists. As infection has developed as a field of study, it became ever more clear that infections could only be understood within the context of the host-pathogen community. This project aims to predict possible drug targets based on host-pathogen interactions rather than just protein-protein interactions within a single organism. Similar to Lee *et al.* (2008) pathogen-host interaction predictions are based on orthology, these interactions are then analysed to identify potential drug targets. This could potentially aid researchers in their continuous battle against malaria and the larger scale battle against pathogen evolution.

To predict *in vitro* host-pathogen interactions DISCOVERY uses an ortholog clustering method called ORTHOMCL. ORTHOMCL is very suitable for ortholog clustering of malaria data for two reasons. Firstly, it is capable of distinguishing between recent paralogs and ancient paralogs, which enables the inclusion of recent paralogs together with orthologs. Secondly, ORTHOMCL was initially developed for the use of malaria data. Identification of *in vitro* interactions is followed by scoring methods to determine the possible *in vivo* interactions that might occur between the *Plasmodium* parasite and the human and mosquito hosts. Scoring measures and weights were applied to 5 different factors to calculate a final score. These final scores allow user input to define the preferred stringency when viewing possible interactions with a single protein. These different factors are sequence similarity, PEXEL/VTS motif presence, microarray expression, metabolic map sharing and sub-cellular locations boundaries.

DISCOVERY'S results and results from two other (Dyer *et al.* and Lee *et al.*) *in silico* prediction methods were compared with Vignali *et al*'s experimental interactions which are based on a yeast two-hybrid approach. Similar to results shown by Doolittle and Gomez these

comparisons had poor results. The next step was to compare the *in silico* results with each other. Dyer *et al's* and Lee *et al*'s results compared poorly with each other. Although DISCOVERY did not compare well with Dyer *et al*'s results, comparisons with Lee *et al.* showed more promise. Poor comparisons with Dyer *et al.* may be due to their unique approach to predict *in vitro* host-pathogen interactions.

This project identified the lack of enough valid and reliable experimental data to evaluate *in silico* prediction methods as a definite challenge for host-pathogen interaction predictors. Although this is a major problem, DISCOVERY improved on older prediction methods with the use of a more applicable ortholog clustering technique and the use of more assessment methods during *in vivo* interaction predictions. DISCOVERY also used scoring methods rather than exclusion methods during the identification of *in vivo* interactions. This allows a user to specify a threshold of sensitivity when viewing interactions.

The true potential of host-pathogen interaction predictions would only be realized when the gap between predictions and evaluation data is bridged.

Table of Contents

TABLE OF FIGURES

TABLE OF TABLES

TABLE OF EQUATIONS

Chapter 1: Introduction

1.1 Malaria

Malaria is one of the world's greatest killers. Infectious diseases like malaria claim millions of lives. Annually 300–500 million clinical cases of malaria are reported, which results in 1.5 – 2.7 million deaths worldwide (Dyer*, et al.*, 2007).

Although malaria is the cause of over a million deaths every year, most of these deaths occur in sub-Saharan Africa. Over 90% of all cases reported occur in sub-Saharan Africa. The cost of prevention and care are worsening conditions even more; the fact that no efficient vaccine for treatment currently exists (Foster and Phillips, 1998) and that acquired parasite resistance has superseded numerous drugs (Kooij*, et al.*, 2006) necessitate urgent attention to malaria research. Consequently, studies on discovering a vaccine or better, less costly prevention methods have become critical.

Malaria is a transmittable disease caused by protozoan parasites from the genus *Plasmodium*. Female *Anopheles* mosquitoes act as vectors or carriers of the *Plasmodium* parasite, which enables cross infection between humans. Approximately two hundred known species of *Plasmodium* exist; about eleven of these species infect humans. Although eleven species infect humans, only five species of *Plasmodium* were included in this project. Four of *(P. vivax*, *P. berghi*, *P. yoelli, P. chabaudi)* of the five cause illness with a low risk of death, while *P. falciparum*, the fifth parasite, is extremely pathogenic and cause progressive illnesses which frequently result in a coma or death.

1.1.1 Lifecycle

Malaria's lifecycle is dependent on two hosts. It starts when the infected carrier host, an *Anopheles* female mosquito penetrates the skin of a human to obtain a blood meal. During penetration, saliva together with elongated sporozoites is inoculated into the bloodstream of a human host. *Via* the bloodstream the sporozoites now travel to the liver, where a process called schizogony (rapid asexual division) takes place. Mature schizonts form during schizogony. These schizonts rupture, to release merozoites, the next lifecycle form. The merozoites may either infect other liver cells or enter the blood stream to invade erythrocytes. Inside the erythrocytes, merozoites develop further, entering either a sexual or an asexual phase.

During the asexual phase a merozoite enlarges within an erythrocyte forming a uni-nucleate ring trophozoite. The ring trophozoites then develop into schizonts with multiple nuclei through mitosis of the nucleus. These schizonts then divide into multiple nucleated merozoites, which cause the erythrocytes to rupture. After the nucleated merozoites exit the erythrocytes, they release toxins into the blood stream which cause fever and chills, and other known symptoms of malaria.

If merozoites enter a sexual phase within the erythrocytes they develop into gametocytes. Gametocytes are able to differentiate into either a male or a female gamete. In contrast with the asexual phase, erythrocytes containing gametocytes do not rupture. They remain in the blood to be extracted by an *Anopheles* mosquito. While the gametocytes are in the mosquito host, they can form male and female gametes, which in turn develop into diploid zygotes. These zygotes differentiate into oocysts which again produces large amounts of sporozoites through mitosis (http://www.dpd.cdc.gov.dpdx).

The cycle repeats itself when the newly developed sporozoites are injected into another human host (Figure 1.1).

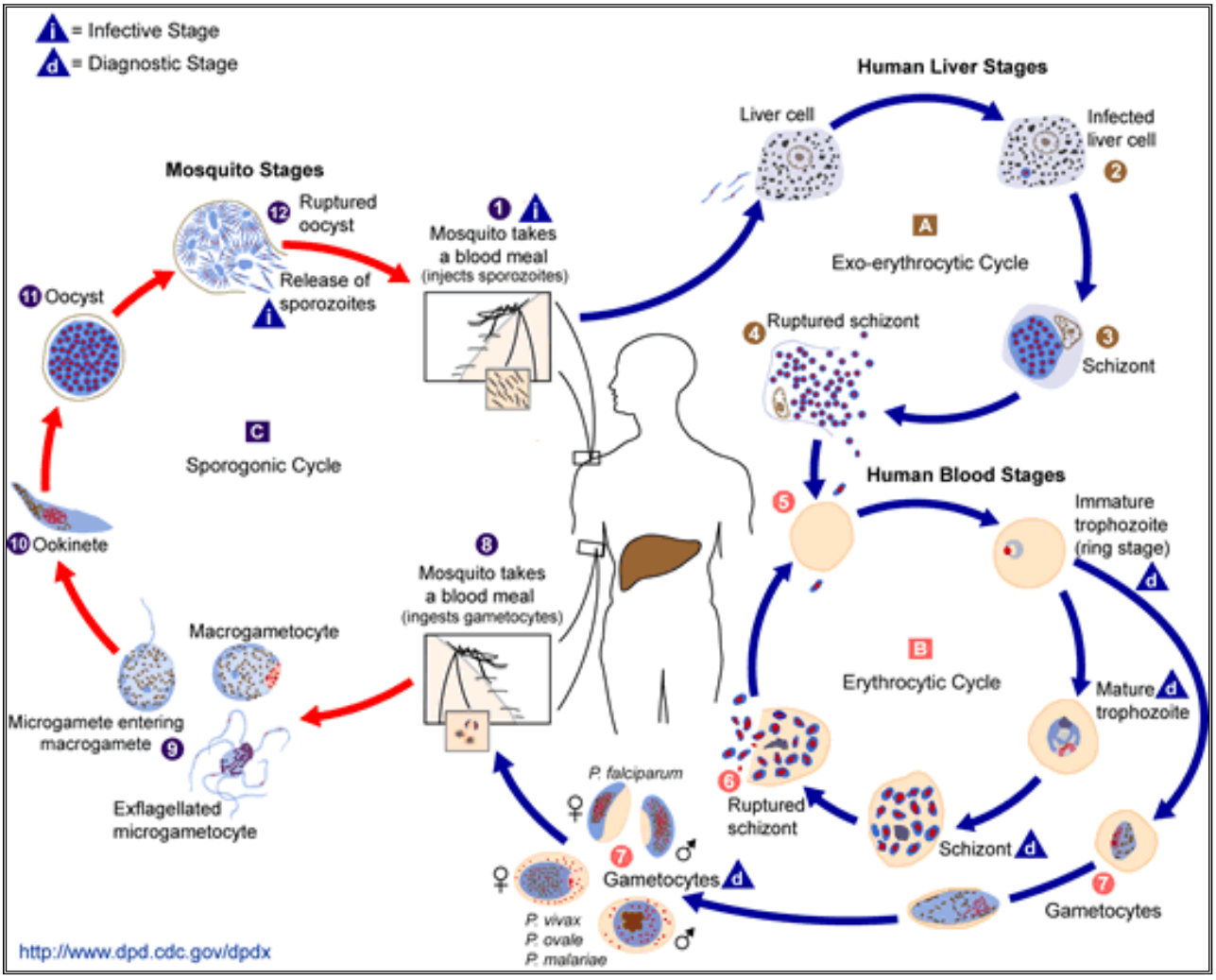

Figure 1.1 A brief diagrammatical summary of the malaria lifecycle
(http://www.dpd.cdc.gov.dpdx)

1.2 Protein-protein interactions (PPI)

The abundance of information gathered during the genomic era made progress of effective genome annotation a critical and inevitable step forward. Annotation allows improved

understanding of genome data and is extremely vital for any genome study. Annotation offers information about the genes that control important functions in organisms. Often genes encode proteins which are responsible for regulating other genes through protein-protein interactions.

Numerous methods have been suggested for studying protein functions and protein-protein interactions. The current research project focused on interactions of proteins between the pathogen and the human host. Since methods of infection are usually initiated by a protein-protein interaction, a few methods of studying protein-protein interactions will now briefly be introduced.

1.2.1 Experimental protein-protein interaction studies

Numerous techniques have been proposed to predict protein-protein interactions. Techniques vary from methods that predict and analyze a single interaction to high-throughput studies that focus on predicting multiple interactions at a time.

1.2.1.1 Single interaction analysis methods

1.2.1.1.1 Nuclear Magnetic Resonance (NMR)

Nuclear magnetic resonance is a prevailing spectroscopic method that provides information about the chemical and structural character of molecules. NMR forms part of a minority of non-destructive methods used for the analysis of structural and molecular dynamics. NMR is based on the behaviour of atoms under the pressure of strong magnetic fields. In biochemistry, these atoms usually are carbon, hydrogen, nitrogen and phosphor.

Atoms that are subjected to a magnetic field behave similar to a compass needle that aligns itself to the earth's magnetic field; the atoms align according to the magnetic field they are exposed to. During NMR a magnetic field is typically introduced to atoms with short bursts of energy on a radio frequency between 40 and 800 MHz. Changing the time between bursts and the strength of the energy bursts, enables the determination of an atom's reaction towards other atoms surrounding it. This way it is possible to determine a molecule's three dimensional structure (O'Connell*, et al.*, 2009).

Presently NMR includes various techniques to predict protein-protein interactions. According to the analysis of the data (April 2009) in the Protein Databank (PDB), 177 of the published protein-protein interactions were determined with NMR (O'Connell, *et al.*, 2009). The three most popular techniques of NMR in PDB are,

- NUCLEAR OVERHAUSER EFFECT (NOE) ~ 70% of NMR interactions
- Residual dipolar coupling (RDC) ~ 15% of NMR interactions
- Pragmatic probes ~ 6% of NMR interactions

As demonstrated above NOE is the most popular technique. NOE is a strongly distance-dependent method based on the interaction between an atom's nucleus with the magnetic dipoles of surrounding nuclei. This dipole-dipole relaxing effect (NOE) is responsible for the distribution of magnetism between nuclei. A combination of NOE with geometric information about bond lengths and angles (from X-ray structures) is normally enough information to determine a molecule's three dimensional conformation.

The growing percentage of approximately 22% of the experimentally predicted interactions in PDB is a good measure of the applicability of NMR. The increase in this percentage could also be

due to the use of newer techniques like RDC and pragmatic probes which seem to become more and more popular.

1.2.1.1.2 Fluorescence resonance energy transfer (FRET) microscopy

FRET is a distance-reliant physical process through which energy is non-radiatively transferred between two fluorescent molecules (fluorophores) *via* long-distance dipole-dipole coupling. FRET produces accurate measurements over distances between 10 Å – 100 Å. If both the donor and acceptor fluorophores fall within the Förster radius FRET is very successful (Förster, 1965). According to studies FRET's efficiency is dependent on the inverse six power of inter-molecular separation, which makes it a sensitive process for the analysis of various biological processes that cause change in molecular proximity (Clegg, 1995; Förster, 1965; Lakowicz, 1988).

FRET microscopy emerged from the combination of FRET and light microscopy. This image-intensity based technique is used to analyze and determine protein-protein interactions in living cells (Periasamy, 2001). FRET's capability of determining protein-protein interactions depends on its ability to detect fluorescent signals from labeled molecules/proteins *in vivo*. During an interaction FRET is activated, diminishing the donor signal as the acceptor signal increases (Herman, 1999).

Numerous other FRET techniques exist, each tuned for specific biological applications with their own advantages and disadvantages. However, all of the FRET intensity methods need processing software to get rid of unnecessary bleed-through components.

FRET is a very exciting novel method of studying protein-protein interactions.

1.2.1.2 High throughput analysis methods

High throughput experimental methods such as yeast two-hybrid (LaCount*, et al.*, 2005) and mass spectometry (Rodland*, et al.*, 2008) had some success in identifying protein-protein interactions. Unfortunately the ability to test a large amount of protein-protein interactions are still lacking. Consequently, alternatives such as computational methods require more attention.

1.2.1.2.1 Mass spectrometry

Mass spectrometry can be described as the world's smallest scale analysis, because of the size of the molecules it can measure. Mass spectrometry can measure molecules like proteins, peptides, carbohydrates, DNA, drugs and other biologically relevant molecules.

Mass spectrometry uses the mass-to-charge (m/z) ratio of a molecule's ions to measure a molecule's mass. Ions are produced *via* forced change of the charge of neutral species. As these ions are formed they are aimed towards the mass analyzer using electrostatic energy. The mass analyzer separates the ions according to their mass-to-charge ratio to enable ion detection. This process enriches scientists with knowledge about the molecular mass as well as structural information about molecules (Rodland, *et al*., 2008).

Mass spectrometry can be characterized through the way molecules are introduced to the mass analyzer. The introduction to the mass analyzer can either be through direct insertion or direct infusion.

1.2.1.2.1.1 Direct insertion

Direct insertion makes use of an insertion plate/probe to introduce sample to the mass spectrometer. Firstly a sample is placed on a probe, this probe is then taken through a vacuum interlock and placed in the ionization field so that mass analysis can take place. The mass spectrometry technique mostly used with direct insertion is called MALDI-MS *(Rodland, et al.*, 2008).

1.2.1.2.1.2 Direct infusion

Direct infusion uses a capillary column that contains the sample either in solution form or in gas form. The advantage of the direct infusion technique is that small molecules can be introduced to the ionization field, without compromising the vacuum interlock. The mass spectrometry technique mostly used with direct infusion is called ESI-MS.

Simplified, mass spectrometry determines a protein's mass before and after any interactions could occur. The mass of the single protein and the mass measured after a possible interaction can then be used to determine if two proteins interact.

1.2.1.2.2 Other methods

Other experimental methods like yeast-2-hybrids, microarrays and other high-troughput screens (Gavin*, et al.*, 2002; Giot*, et al.*, 2003; Ho*, et al.*, 2002; Ito*, et al.*, 2001; Ito*, et al.*, 2000; LaCount, *et al*., 2005; Li*, et al.*, 2004; Rual*, et al.*, 2005; Stelzl*, et al.*, 2005) have also been introduced to identify protein-protein interactions.

Since these methods also have a potential for determining host-pathogen protein-protein interactions, they are discussed in more detail in section 1.3.2.1.

1.2.2 *In silico* protein-protein interactions studies

In silico level protein-protein interactions are usually only predicted within a single organism. These prediction methods use different techniques, from sequence signature pairs (Sprinzak and Margalit, 2001) to protein-domain profiles (Kim*, et al.*, 2002; Ng*, et al.*, 2003) and sequence homology (Yu*, et al.*, 2004).

As pathogen-host infections (interactions) have developed as a field of study, it became ever more clear that infection is only really understood in the context of the host-pathogen community (Jackson*, et al.*, 2006). Given this key insight it is essential to take into account pathogen evolution studies. Pathogen evolution is one of the greatest hurdles to overcome in the development of effective disease control. The evolution of a pathogen leads towards drug resistance and an increase of pathogen virulence (Grech*, et al.*, 2006). This makes drug discovery a tedious, but an incredibly important task.

1.3 Host-pathogen interactions

1.3.1 Origin of host-pathogen interactions

Most of the terms used to explain host-pathogen interactions have existed for approximately a century (Casadevall and Pirofski, 1999; Casadevall and Pirofski, 2000). Initially microbes were seen as the invader that causes disease. Further studies on the characteristics of microbes revealed that pathogen-host interactions do not always result in negative effects or disease. This meant that not all microbes were pathogens. Attention was shifted to the identification of harmless microbes and the definition of the different circumstances in which microbes exist without causing disease.

Terms like commensal and opportunist were suggested for describing this strange occurrence between microbes and hosts. These terms initially originated to describe microbe characteristics, rather than host-pathogen interactions. Thus, it became important to reconsider the definition of each term (Casadevall and Pirofski, 2001). Subsequently, studies developed towards a holistic perspective which includes both host and pathogen characteristics in a framework for studying host-pathogen interactions. A summary of these steps now follows.

Since the germ line theory was accepted it was believed that microbes could be classified according to Koch's postulate; if a certain microbe passed all conditions it was defined as a pathogen. It soon became obvious that many microbes existed and that only a few were responsible for disease in human hosts (Zinsser, 1914). Classification according to Koch's postulates led to harmless microbes being classified as pathogens; and pathogens as harmless microbes. This guided a movement towards the formulation of new ideas and terminology.

At the beginning of the 20th century it became clearer that pathogenicity was not a stable or consistent definition of microbes, because pathogens did not always cause disease. A key finding in laboratory studies provided the answer, it was found that a host could increase or decrease the virulence of a pathogen. This meant that it was possible that a host could influence pathogens' ability to cause disease. Development of vaccines originated from this discovery. Later studies on infection identified some hosts as carriers of pathogens, but not as carriers of diseases (Henrici, 1934). This led to the hypothesis that certain hosts were more susceptible to disease than others. Although, the carrier state of hosts was not understood (Smith, 1995) it opened a door for defining host-microbe interactions as an integrated whole and made researchers more aware of the influence of microbe-host interactions on infection. This awareness supported a movement away from the microbial perspective and towards an integrated perspective that includes both microbe and host factors during infection. Although

most of the terms on microbe-host interactions are adequate, they are based on a microbial pathogenesis framework and do not take into account the variable conditions of microbe-host interactions (Casadevall and Pirofski, 1999).

There seems to be a lot of uncertainty around the theme of pathogen-host interactions. It is therefore necessary to recognize the need to study pathogen-host interactions as an integrated whole. Only then can infection be truly understood.

More knowledge about the mechanism of infection may guide effective drug discovery and development of new vaccines. This project addresses host-pathogen interaction predictions according to the holistic approach described above. Each aspect of the holistic approach is discussed in the following section.

1.3.1.1 Pathogens

A pathogen is defined as a microbe capable of causing host damage (Casadevall and Pirofski, 1999). The definition of a pathogen has been a problem for many years, because it has been studied from a solely microbial perspective. It has recently been recognized that a pathogen can only be defined in respect to a specific host. Viewing pathogens from an integrated host-microbe framework already clarified a lot of uncertainties about pathogen attributes and adaptation between a host and a microbe. Although the need for a new base framework was recognized, the lack of knowledge about microbe-host integration studies haltered development towards this new framework. Researchers are continuously trying to find new and better ways to handle this challenge (Blaschke*, et al.*, 2001; He*, et al.*, 2009; Krallinger and Valencia, 2005; Zhou and He, 2008).

Hence, defining a pathogen and its attributes still remains a problem. Studies about pathogens appear to have identified three main criteria for classification of pathogens.

- Severity of induced disease
- Route(s) of infection
- Virulence and infectivity

Most pathogens are dependent on a host to be able to persist and survive. The understanding of the general lifecycle of pathogens is therefore of extreme value. The lifecycle of a pathogen usually includes the following processes (Figure 1.2). Firstly, entrance into a host through anatomical access, this includes passing through the natural barriers of a host. This is followed by spreading of disease, the escape from host defense and the persistence of the pathogen within the host body. Successful persistence is obtained through self-regulation and evolution towards a balance between pathogen virulence and the ability to evade host defense systems. Finally, pathogens must have mechanisms to penetrate or damage host cells, organs and the rest of the body (Walker*, et al.*, 2003). Damaging host cells is important for the pathogen; it is used for self-defense against the host and to obtain necessary resources for survival (Walker, *et al.,* 2003). This whole cycle starts again when transmission between hosts occurs.

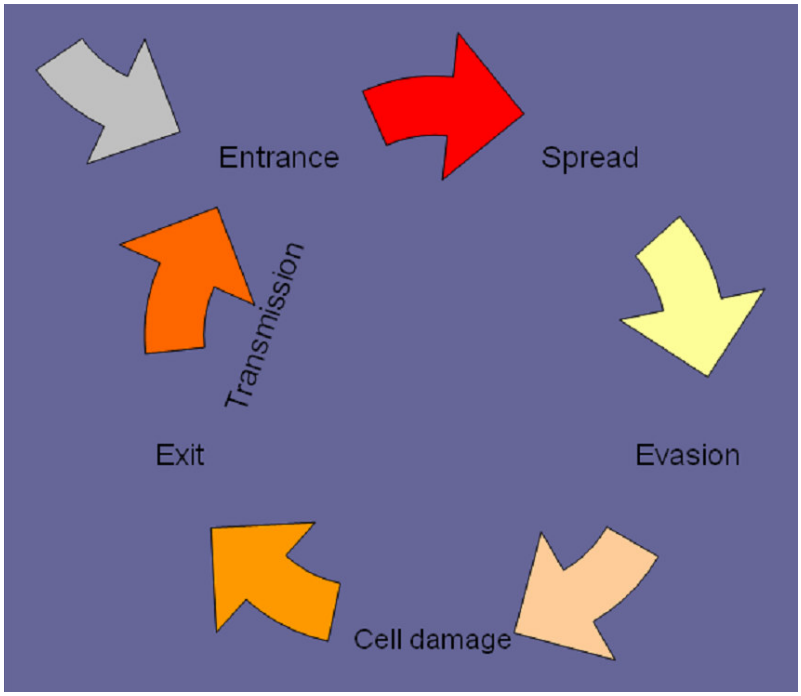

Figure 1.2 General pathogen life cycle

Mechanisms of transmission are greatly dependent on the ecology of a pathogen. Consequently respiratory pathogens are very likely to be airborne pathogens, while pathogens that infect the intestinal tract usually exist in water and foods. Pathogen transmission is classified into two general groups of contact; a direct transmission and an indirect transmission group (Caraco and Wang, 2008). Each group has various mechanisms of transmission. Direct transmission happens when a pathogen is transmitted directly from one host to another. Pathogens involved with direct transmission are typically extremely sensitive to their environment and thus incapable of surviving outside a host for long periods of time. An example of direct transmission is sexually transmitted diseases (STD), where transmission takes place by the transfer of blood, saliva, semen and other bodily fluids between hosts. Indirect transmission takes place when an agent is needed to transmit a pathogen from an infected host to a susceptible host. An agent may be animate or inanimate. Typical animate agents are wind, water, food or mosquitoes; inanimate agents are non-living materials like cloths, toys and surgical tools. An example of indirect transmission is malaria, where *Plasmodium falciparum* is transmitted through an agent, female *Anopheles gambiae* mosquitoes, to infect human hosts.

1.3.1.2 Hosts

Walter Bradford Cannon (Cannon, 2009) defined homeostasis as the property of a system, either open or closed, that regulates its internal environment and tends to maintain a stable, constant condition.

As time passes homeostasis adapts in many ways to keep equilibrium. One such adaptation is the internal development of mechanisms to protect a host from harmful microbes. The body's internal self defense mechanisms are collectively called the immune system. The immune system differs between animals and plants.

1.3.1.3 Animal immune response

The animal immune response is a quick yet accurate reaction against microbial pathogens. The immune response primarily exists as two defense responses called the innate and adaptive responses.

1.3.1.4 Innate response

The innate response (Accolla, 2006; Modlin and Doherty, 2003) is the fast expression of germ line–encoded pattern-recognition receptors on macrophages, dendritic cells (DCs), NK cells and epithelial cells. These receptors recognize specific biochemical patterns in the membrane presenting molecules of foreign objects or pathogens. Thus, the receptors enable interaction with pathogens. In this way the above mentioned cells could attack pathogens. The activation of an innate response can either lead directly to antimicrobial pathways or indirectly via cytokines which may regulate the adaptive response.

1.3.1.5 Adaptive response

The adaptive response (Accolla, 2006; Modlin and Doherty, 2003) on the other hand is a slow reaction where highly specific gene rearrangement of T-cell and B-cell receptors takes place. This receptor rearrangement allows for more accurate foreign cell recognition, which in turn enables innate response cells to attack and interact with the foreign / pathogen cells.

1.3.1.6 Plant defence system

Similar to animals, plants (Szarka*, et al.*, 2002) also have a general defense response and a more specific defense response.

1.3.1.7 General defence system

The general defense system (GDS) is a fundamental attribute of plants which protects plants against biotic as well as abiotic substances. The general defense reaction is based on the enlargement and division of cells influenced by stress. This enlargement and division causes the compacting of the tissue of a cell.

The general defense system is not specific or limited to only one pathogen species. A reaction is recognized by an initiation at a low stimulus threshold, which causes a quick reaction to strengthen and protect the stress-influenced cells. Although most plants have a general defense system, general defense systems may vary in effectiveness.

The general defense system is the plants' first line of defense, without it continual protection against pathogens cannot be expected. Studies on the general defense system in injured cells, show no defense response against pathogens which might mean that the general defense system only has a preventative function.

1.3.1.8 Specific defence system

The specific defense system is the reaction of cells that are attacked or infected by pathogens. A specific defense reaction time is noticeably longer than that of the general defense response and also has a higher threshold stimulus. Under normal circumstances, this difference in threshold causes a general defense reaction followed by a specific defense reaction (Szarka, et al., 2002).

Thus, it is assumed that the general defense system has the role of immune response and that the specific defense system only corrects the general defense reaction's shortcomings in plant disease resistance.

1.3.2 Methods of studying host-pathogen interactions

The abundance of information gathered during the genomic era made genome annotation a critical and inevitable next step. Annotation converts this information into useful knowledge about genomes, which is extremely vital for characterizing disease or infection. Annotation offers information about the proteins that control important functions in organisms, allowing studies about the functionality of proteins.

Numerous methods have been suggested for studying protein functions and protein-protein interactions.

1.3.2.1 Experimental methods for analyzing host-pathogen interactions

Currently laboratory experiments on protein-protein interactions vary from expression profile methods like SAGE (Velculescu*, et al.*, 1995), differential display (Wang and Feuerstein, 1997) and cDNA microarrays (Hegde, *et al.*, 2000) to hybrid techniques like yeast two-hybrid

screening (LaCount, *et al*., 2005) and molecular techniques like mass spectrometry (Rodland, et al., 2008).

1.3.2.1.1 Expression profiles

A genomic correlation exists between expression profiles and protein-protein interactions. Studies analyzed this correlation to be stronger than a correlation between expression and random proteins (Bhardwaj and Lu, 2005; Grigoriev, 2001). This revealed protein-protein interactions reflected in expression profiles.

Expression-based techniques aim to study the measurements of mRNA expression levels of identical cells, either in the same cellular tissue or different cellular tissue. Expression varies between different environmental conditions or different developmental stages. The difference in expression indicates the level of gene activity during infection. Gene activity is used to identify proteins of hosts and pathogens that may possibly interact.

The use of expression profile-based methods to predict protein-protein interactions, within a single species and between different species, is becoming more reliable with the increase of high quality interaction data.

1.3.2.1.1.1 Microarrays

Microarray expression analysis has a few essential features that makes it a popular method to use for mRNA expression profiling (Hegde, *et al.*, 2000). DNA segment sequences or oligo sequences representing the genes are amplified via PCR and then mechanically spotted onto a microscope glass slide or Affymetrix slide at high density. The XYZ robotic system enables

simple, straightforward manufacturing of microarrays slides containing the total set of genes of a microbial genome or multiple eukaryotic cDNA clones.

After creation of microarray slides, labeled probes are constructed. These labeled probes are also called co-hybridized assays. Probes are labeled through mRNA modification with either one of two fluorescent dye labels (Shalon*, et al.*, 1996).

Probes are hybridized onto slides to determine the expression levels of target sequences. The expression levels are calculated based on the ratio by which a probe hybridizes with any given target. Confocal fluorescent scanning is used for the measuring of the fluorescent intensities of molecular bonds in an array.

Successful expression analysis *via* microarrays depends on the development and successful execution of various laboratory techniques for fluorescent intensity normalization. These techniques can be divided into three steps.

- Array manufacturing
- Probe preparation and hybridization
- Data collection, normalization and analysis

The successful implementation of these techniques allows the identification of co-expressed genes, which aids in predicting protein–protein interactions.

1.3.2.1.1.2 Serial analysis of gene expression

Another expression profile technique is called serial analysis of gene expression (SAGE) (van Ruissen and Baas, 2007). SAGE is a profiling method that has the ability to analyze thousands of transcripts simultaneously on qualitative and quantitative levels (Velculescu*, et al.*, 2000). SAGE can be broken up into several key steps (Figure 1.3).

The first step of SAGE would be to isolate the mRNA (gene data) from cells. These mRNA molecules are captured by oligo-dT beads. The beads exploit the characteristic of mRNA which always contains a poly-A tail. Each bead contains many poly-T tails, with each tail acting as a magnet for a single mRNA sequence. Researchers discovered that fourteen nucleotides are enough to match mRNA to its complementary gene. This simplified things considerably since tags rather than whole RNA strands could be used to identify unique transcripts. Each mRNA strand is cut by various enzymes. Firstly enzymes cut off the “sticky ends” of the RNA strain and then a primer for another enzyme is attached. The next enzyme binds to the primer and cuts the original RNA to form a 14-mer tag.

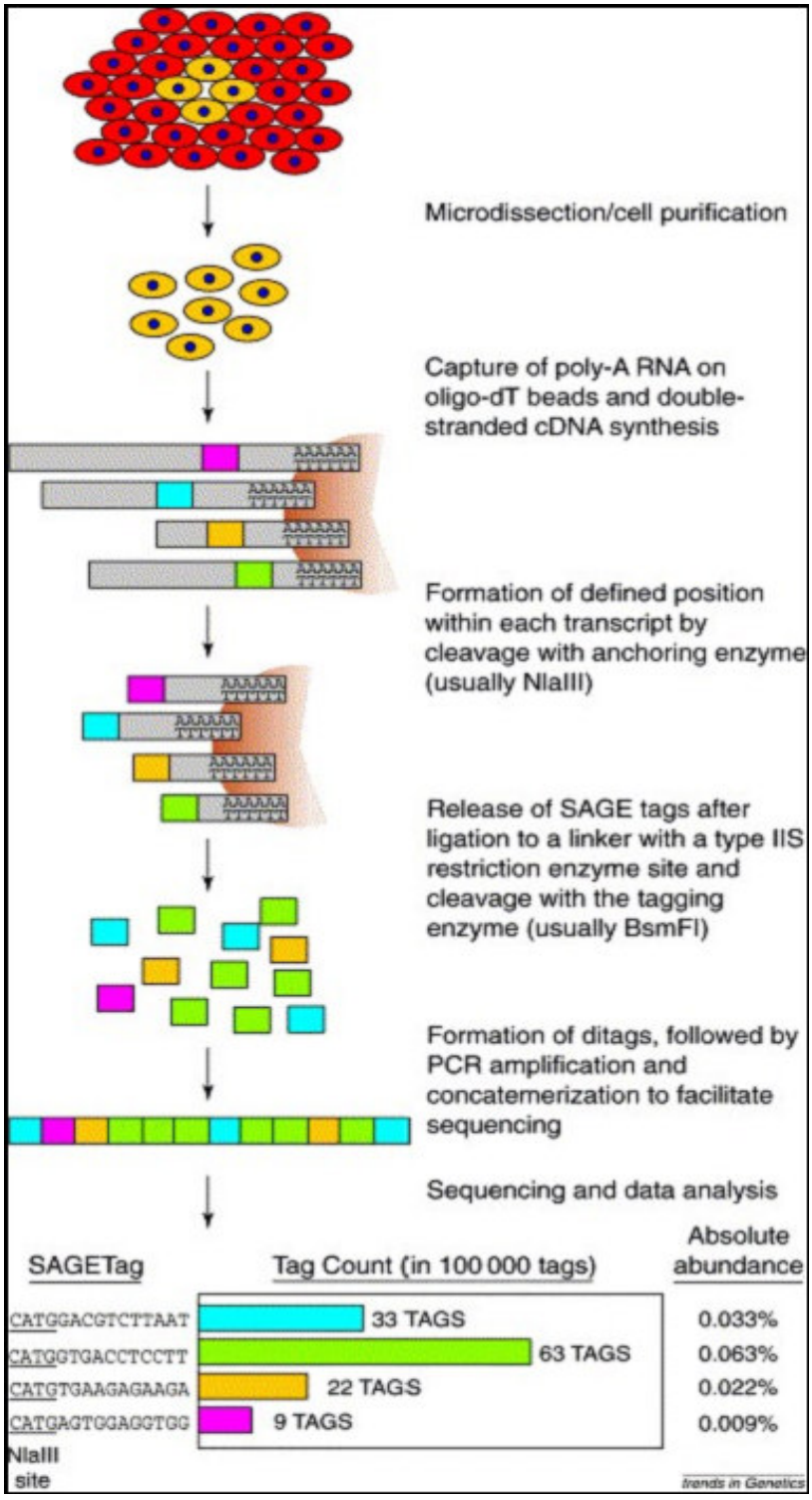

Figure 1.3 Steps of SAGE analysis
(Velculescu, *et al.*, 1995)

All the 14-mer tags are ligated to one another to form a long concatemer. The concatemer is then cloned (PCR) and sequenced. Consequently the frequency of the tags discovered after sequencing represents the level of transcripts present.

A major drawback of SAGE is that it needs a relatively high amount of RNA to analyze differential expression. Recent modifications to resolve this problem, resulted in MICROSAGE (Datson*, et al.*, 1999), a new technique that needs 500 – 5000 fold less RNA starting material. Several other modifications to SAGE have also been proposed, LongSAGE (Saha*, et al.*, 2002), RL-SAGE (Gowda*, et al.*, 2004) and SuperSAGE (Matsumura*, et al.*, 2008) are some of the most recent. Most of these modifications focus on techniques to capture longer tags, which increases the confidence in results.

The basic method of SAGE is based on the fact that not all genes are employed within a cell and that gene activation differs for different kinds of cells. SAGE measures gene activity (mRNA levels) to construct cell specific profiles of normal functioning cells. These profiles can then be used to compare normal cell activity against diseased cell activity and therefore identify proteins that are active during infections as possible interacting proteins.

1.3.2.1.2 Other methods

1.3.2.1.2.1 Mass spectrometry

A summary of the basics of mass spectrometry has been discussed as method identifying protein-protein interactions (see Section 1.2.1.2.1). The fast development of the mass spectrometry technique has made it possible to use mass spectrometry to determine possible host-pathogen protein-protein interactions.

Rodland et al. (2008) mentions a few techniques of mass spectrometry to predict host-pathogen interactions. One of these exciting methods aims to identify interactions *in vivo*. This mass spectrometry approach uses chemical cross-links between interacting partners within living cells. Cross-linkers are able to form covalent bonds between interacting proteins. The strength of cross-linker bonds with each other can vary from non-specific to brief and even highly specific bonds.

The advantage of the applying cross-links before cell lysis is that it allows the identification of brief interactions *in vivo* (Rodland, *et al.*, 2008). These brief interactions are interesting and elusive, because they usually occur via signals and are important for cellular processes.

Currently mass spectrometry methods already have a huge impact on the identification of protein-protein interactions within the same species and between different species *in vivo* and *in vitro*. The future expansion of this technique holds a lot of promise for protein-protein interactions in general.

1.3.2.1.2.2 Two-hybrid analysis

As with microarrays, the two-hybrid analysis needs prior knowledge. One limitation of the two-hybrid system is that it only focuses on testing for interactions between a single known protein or DNA sequence and multiple other sequences.

The two-hybrid assay is a molecular biology technique that allows for the discovery and testing of protein interactions with either proteins or DNA sequences (Hurt*, et al.*, 2003; Joung*, et al.*, 2000; LaCount, *et al*., 2005). The process is based on physical binding between proteins and

DNA sequences or other proteins. The central idea of the two-hybrid system builds on the behavior of eukaryotic transcription factors.

Eukaryotic transcription factors usually are modular, consisting of 2 domains; a binding domain (BD) and an activation domain (AD) (Hurt, *et al*., 2003; LaCount, *et al*., 2005). Characteristic to eukaryotic transcription factors, their domains do not need to be in direct contact with each other to be able to activate transcription (Verschure*, et al.*, 2006). The two-hybrid system exploits this attribute by indirectly binding the transcription factor domains through an interaction of a single known sequence with other possible sequences to enable the transcription of a reporter gene. If the reporter gene is transcribed it implies that a new interaction is discovered, which needs further analysis.

1.3.2.1.2.2.1. Yeast two-hybrid (Y2H)

The best known two hybrid technique is called the yeast two-hybrid system (Gietz*, et al.*, 1997). This system uses a genetically modified strain of yeast which lacks certain nutrients for biosynthesis. Further strain modifications allows for the introduction of foreign DNA to the cells in the form of a plasmid.

Two different kinds of plasmids are used. The one plasmid is constructed so that the product it produces contains a protein merged with a DNA binding domain and the other plasmid so that a protein is merged with a DNA activation domain.

The protein merged to the DNA binding domain is referred to as the bait. The bait protein is usually a single known protein subjected to possible bonds from several other proteins or DNA sequences. These other proteins or DNA sequences are referred to as the prey proteins. Each

prey protein is merged to a respective activation domain. During yeast-2-hybrid screening the separate bait and prey plasmids are simultaneously introduced to the modified yeast strain.

If a bait protein interacts with a prey protein, the binding domain and the activation domain of the transcription factor are indirectly linked and ready to initiate transcription (Joung, 2001). The binding domain then binds to the upstream activation sequence and transcription can start (Figure 1.4).

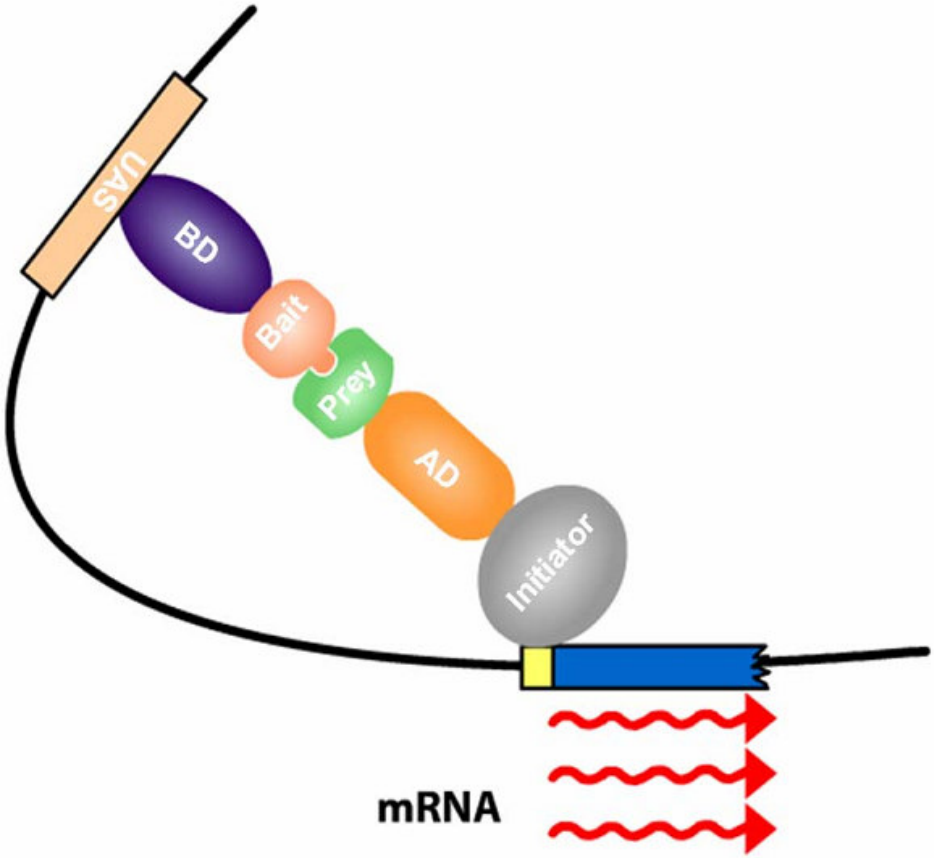

Figure 1.4 Yeast two-hybrid activation complex
(http://pustaka.ictsleman.net/biologi/biologi_flash/58_yeasttwohybrid%5B1%5D.swf)

During transcription some kind of a reporter gene is transcribed which signals that an interaction occurred. In the case of the lack of nutrients for biosynthesis, the reporter gene would transcribe the nutrients needed for biosynthesis and cell growth will be the signal to an interaction. If cell growth appears, the prey proteins are isolated from the yeast cells for further analysis. Cell death would be the signal that no interaction was identified.

The yeast two-hybrid approach is currently one of most promising methods used to detect and analyze protein-protein interactions. Unfortunately, two-hybrid results are prone to a significant ratio of false positives (interactions wrongly detected as true) and false negatives (true interactions not detected). While measurement of false positives is difficult to determine, false negative rates can vary from 70% - 90% (Stellberger*, et al.*).

1.3.2.1.2.2.2. Known yeast -2-hybrid studies in malaria

LaCount *et al*. successfully applied Y2H to predict physical interactions between *H. sapiens* and *P. falciparum*. Their results identify 1308 proteins involved in 2846 unique protein-protein interactions obtained from 32000 yeast two-hybrid screens. Their analysis shows that 82% of all predicted physical interactions contain at least 1 unknown protein, while 33% includes interactions with 2 unknown proteins.

All yeast two-hybrid predicted interactions that had a higher affinity than interactions occurring by chance were selected and further analyzed. Clusters of functionally related proteins could be derived from these interactions. The clusters potentially provide important insights into the life cycle of *Plasmodium* within the human host.

Further analysis of the yeast two-hybrid interactions from LaCount *et al*. includes comparison analysis *via* the NetworkBLAST tool (LaCount, *et al*., 2005) and inference of functionality using PlasmoMAP (Date and Stoeckert, 2006).

The analyses of an additional 10000 searches for human activation domain libraries are currently commencing (LaCount, *et al*., 2005).

1.3.2.2 *In silico* prediction of host-pathogen interactions

The main characteristic of a host-pathogen system is the mechanism through which organisms interact. The most general mechanism is protein-protein interactions, where a pathogen's proteins target host proteins.

The development of computational methods to identify protein-protein interactions between a pathogen and a host will provide enormous advantages in identifying possible drug targets. Unfortunately sources of protein-protein interactions are limited. Although several protein-protein interaction prediction programs exist, they focus on a single organism viewpoint. Protein-protein interaction prediction programs with the holistic viewpoint of both a host and a pathogen are very scarce.

In the case of malaria, two previous computational studies for predicting host-pathogen interactions have been performed (Dyer *et al*. and Lee *et al*.). While they are briefly introduced here, they will be discussed in extensive detail in the following chapters.

1.3.2.2.1 Computational prediction and Bayesian statistics

Dyer *et al*. integrated data from publicly available intra-species databases with protein-protein interaction profiles to build a framework able to predict protein-protein interactions (Dyer, *et al., 2007)*. To increase confidence in predictions, Bayesian statistics are used to calculate reliability of the prediction of a protein to interact with two specific domains.

Firstly, host and guest proteins are grouped and classified either as two host proteins that interact with a pathogen protein (H-H-P) or two pathogen proteins that interact with one host protein (H-P-P). For each group the domain distances are calculated over all the distribution distances. Then microarray datasets of the different phases of the parasite's lifecycle are collected. This is followed by calculating Spearman's correlation between the group distance calculated earlier and expression profiles. Finally, these groups are enriched with functionality Gene Ontology functions (Ashburner*, et al.*, 2000) for pathogen and host proteins in a predicted interaction.

Dyer *et al*. also recognized the need for further filtering of data to minimize noise. To solve this problem they proposed filtering on organelle level (Dyer, *et al.*, 2007).

1.3.2.2.2 Ortholog based approach

Lee *et al*. (2008) introduced a host-pathogen protein-protein interaction predictor based on orthologs. Lee *et al.* used the POINT interaction database in combination with orthologs from the HOMOLOGENE database to predict possible host-pathogen interactions. These host-pathogen interactions are predicted according to the fact that the predicted interactions shared ortholog groups between the proteins in the POINT interactions and the host and pathogen proteins from HOMOLOGENE.

More accurate predictions were obtained by choosing only the interactions that had high scores within the POINT database. The article also introduced further filtering according to sharing of genome ontology. They identified that more host-pathogen interactions occurred between proteins involved in metabolic and cellular processes. After genome ontology filtering, interactions were filtered according to translocation signals which led to the prediction of 95 interactions between *H. sapiens* and *P. falciparum*.

DISCOVERY'S host-pathogen protein-protein interaction prediction technique utilizes a technique more similar to the work of Lee *et al*, which is also based on orthologs.

1.3.2.2.3 Host pathogen interaction databases

Finaly, there are also some online databases that make pathogen-host interactions readily available, the data from these databases are usually based on experimental data or highly currated literature searches. Some examples of these are PHI-BASE (Winnenburg*, et al.*, 2006) and PHIDIAS (Xiang*, et al.*, 2007). Although containing useful information, they are quite limited in scope and cover only the genomes of a few species.

1.4 Problem Statement

Currently the amount of annotated data available for malaria severely limits studies on host-pathogen protein-protein interaction predictions. Only a few such studies exist (Dyer, *et al.*, 2007; Lee, *et al.*, 2008). The most popular method to determine the function of an unknown sequence is to determine the similarity (homology / orthology) of the sequence with regard to annotated proteins or genes. Numerous protein-protein interaction prediction methods use homologs / orthologs to overcome this drawback of limited annotated data (O'Connell, *et al*., 2009; Yu, *et al*., 2004).

Most protein-protein interactions studies are focused on one species, where protein-protein interactions predictions are integrated into interaction networks to get a better understanding of a species' regulatory mechanisms and protein functions. This approach of focusing on only one species has also been applied to identify protein-protein interactions between species based on the assumption that disease or infection is caused by only pathogen features. This assumption continuously revealed discrepancies where a pathogen did not always cause disease. Recently this led to the realization that both host features and pathogen features should play a role during infection, meaning that host-pathogen protein-protein interactions or the lack thereof results in successful infection of a host and the cause of disease.

Host-pathogen protein-protein interaction predictions reveals the criteria for defining novel drug targets and is therefore of utmost importance to researchers involved in lead design and screening projects for malaria. This project aims to predict host-pathogen protein-protein interactions in malaria, taking both host and pathogen features into account. It forms part of a larger project entitled DISCOVERY (Joubert*, et al.*, 2009). The DISCOVERY project provides researchers with a resource for the selection of putative target proteins and lead ligand molecules for the malaria parasite. This study aims to improve the annotation of possible host-pathogen interactions in malaria, with a view to the optimization of the drug discovery process.

1.5 Specific Aims

To enable host-pathogen protein-protein interaction predictions, specific aims included:

1. Integrate the annotated gene products of *P. falciparum, P. vivax, P. berghei, P. chaubaudi, P. yoelii, H. sapiens* and *A. gambiae* with the data from the publicly available DIP and MINT protein-protein interaction databases into a *fasta* file.

2. Use the *fasta* file to generate custom ortholog clusters from previously integrated data using OrthoMcl.
3. Predict possible *in vitro* host-pathogen protein-protein interactions using the annotated data from the 7 species in Discovery, interaction data from Dip and Mint and the ortholog cluster data generated by OrthoMcl.
4. Construct a scoring system that measures an interaction's ability to occur *in vivo* according to sequence similarity, sub-cellular locations, Pexel / Vts motif presence, metabolic maps sharing and microarray expression results.
5. Specify the default threshold score for identifying *in vivo* interactions
6. Compare Discovery's prediction results with a *H. sapiens* - *P. falciparum* empirical dataset and two known *in silico* predictors and check these predictions against Negatome data.
7. Integrate the host-pathogen protein-protein interactions results into Discovery's web server.

Chapter 2 discusses the implementation of host-pathogen interactions in Discovery, while Chapter 3 compares Discovery's host-pathogen interaction predictions with other methods and Chapter 4 provides a concluding discussion.

Chapter 2: The design and implementation of host-pathogen interaction predictions

2.1 Introduction

After the genomic era, computational methods have become increasingly important for annotating genomes. The greatest challenge now, is to successfully exploit the knowledge generated during the genomic era. The rationale behind the annotation, analyses and integration of genome data is to implement computational methods to filter useful information.

Currently, a reasonable amount of *in silico* protein-protein prediction tools exists. These tools predict proteins and protein families, protein structural features and structures, protein location, protein functionality and even interactions between proteins. Storage of experimentally proven data in universal databases such as UNIPROT (Apweiler*, et al.*, 2004) and GENBANK (Benson*, et al.*, 2009) has also proven to be of great advantage.

This study forms part of a greater project, entitled DISCOVERY. The DISCOVERY project will now be discussed in some detail, before the approach used in predicting host-pathogen interactions specifically is addressed.

2.1.1 Discovery

DISCOVERY (Joubert, *et al.*, 2009) focuses on the annotation of properties relative to the drug design process in malaria. Currently, DISCOVERY contains annotations of *Plasmodium falciparum*, *Plasmodium vivax, Plasmodium berghei* and *Plasmodium yoelii* from PLASMODB (Aurrecoechea*,*

et al., 2009) and annotations of *Homo sapiens* and *Anopheles gambiae* from ENSEMBL (Hubbard, *et al.*, 2002). This data is further integrated with information from DRUGBANK, KEGG and PDB-ligand, DIP and MINT. Protein functionality was predicted using INTERPRO. The integrated results of these sources gives DISCOVERY the ability to predict features like protein domains, motifs, metabolic processes, orthology, protein-ligand interactions, protein-protein interactions and host-pathogen interactions (Figure 2.1). DISCOVERY illustrates the power of data integration in pursuing more knowledge about a genome.

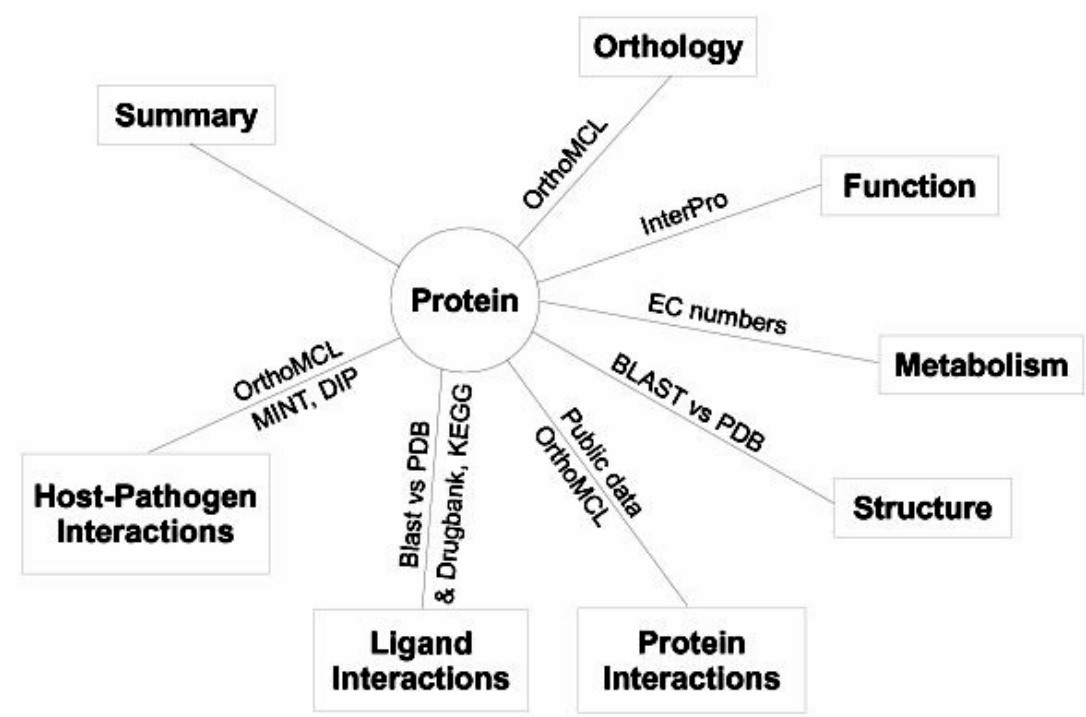

Figure 2.1 The protein features that DISCOVERY extrapolates from data integration
(Joubert, *et al.*, 2009)

DISCOVERY can be accessed at the publically available webpage http://malport.bi.up.ac.za:8150. DISCOVERY allows four ways of mining data; these methods can be seen on DISCOVERY's homepage (Figure 2.2).

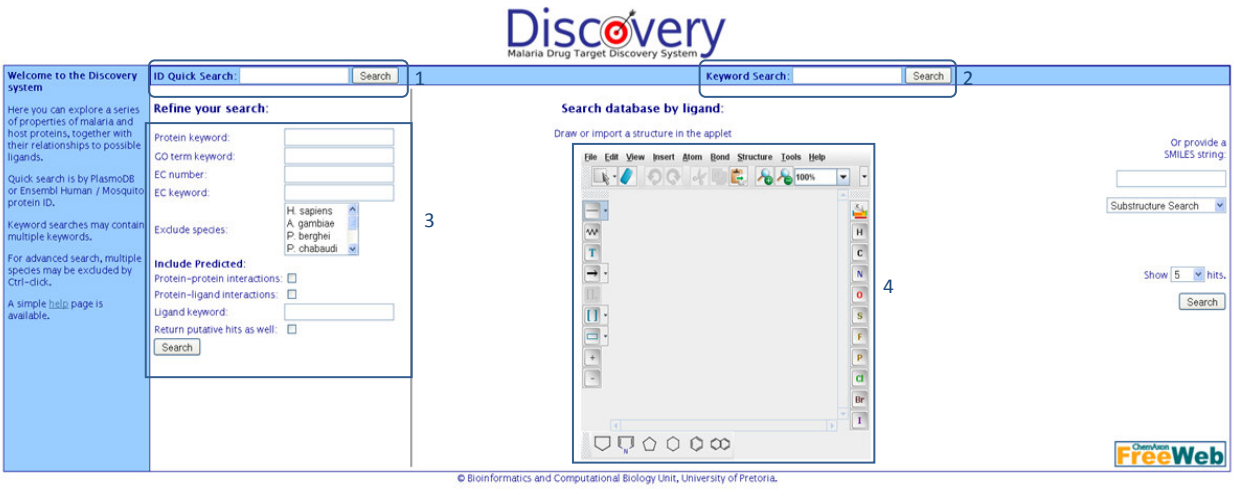

Figure 2.2 DISCOVERY's homepage allows four different methods to mine data about malaria

2.1.1.1 Quick search

In figure 2.2, the first access method is a quick search which allows a user to query information about malaria using a protein identity (id) from PLASMODB or ENSEMBL. If the queried protein was not found within DISCOVERY's database the user is informed that the specified protein does not exist. A successful search will result in DISCOVERY returning all the information available on the queried protein, in a tab-format (sample output to follow). For illustration purposes a sample search on **PFF0940c** was performed.

The first tab is the **Summary** tab which contains all the basic information about the protein PFF0940c, such as protein names, descriptions, functions, sequence, GO ontologies and EC numbers. From this tab it can be seen that **PFF0940c** is described as a cell division cycle protein.

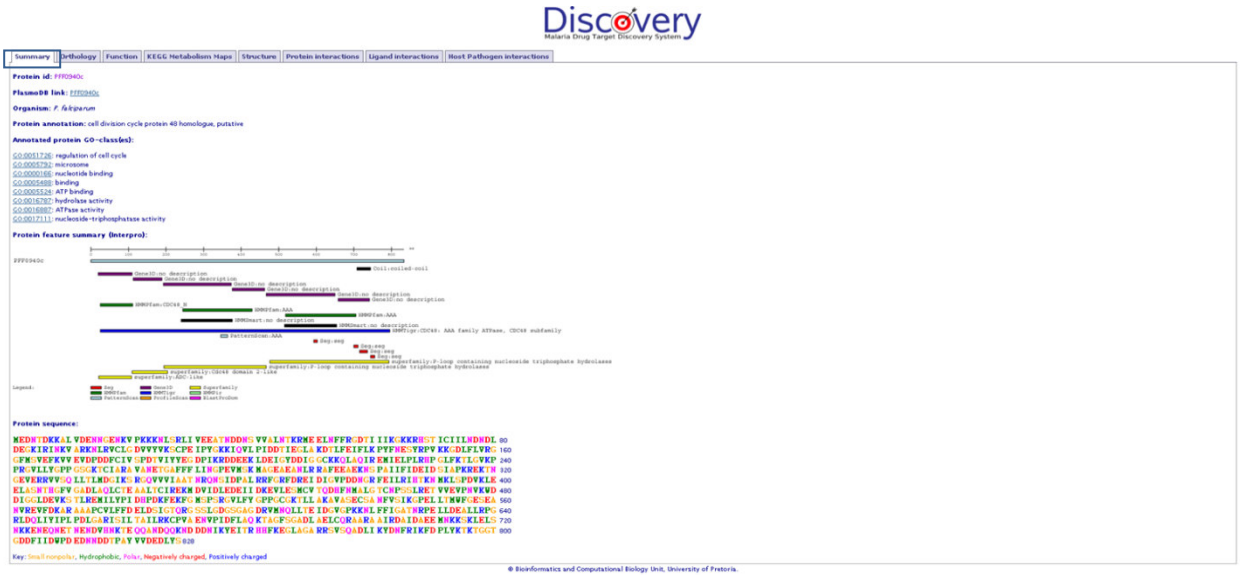

Figure 2.3 Summary tab of DISCOVERY'S quick search results

The next tab is called the **Orthology** tab. This tab contains details about ORTHOMCL results; a T-COFFEE alignment was done on all the orthologs identified for **PFF0940c**. In this figure it can be seen that the ORTHOMCL results show other proteins that also has the description, cell division cycle protein, as orthologs of **PFF0940c** (figure 2.4).

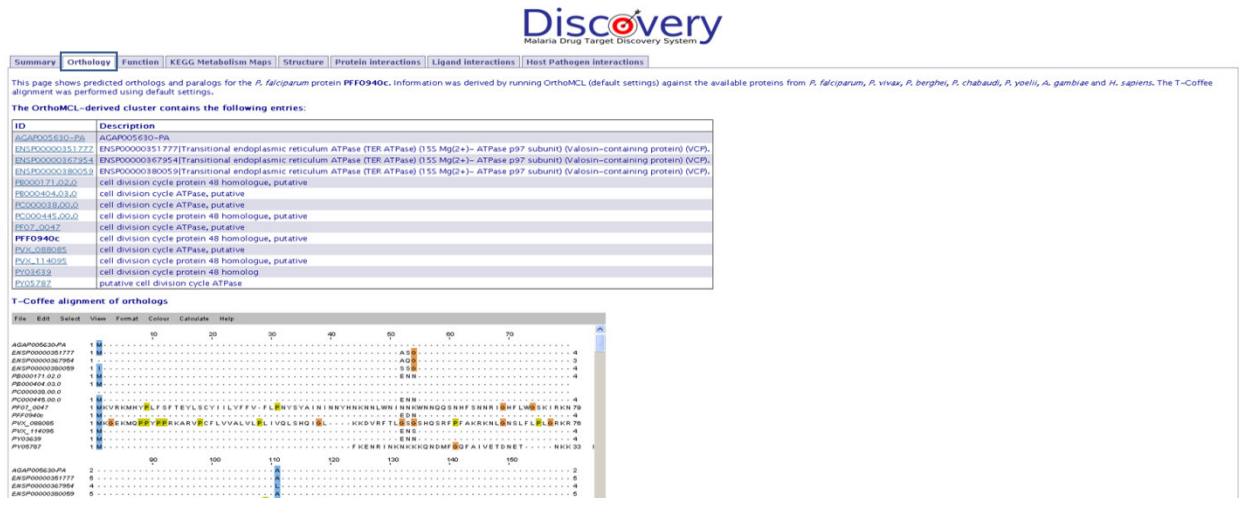

Figure 2.4 Orthology tab, containing ORTHOMCL results alligned with T-COFFEE

The **Function** tab provides a tabular summary of INTERPRO hits, together with a graphical summary of functional domains predicted by INTERPRO SCAN. The tabular summary also includes short descriptions of the domains, together with respective confidence scores (figure 2.5).

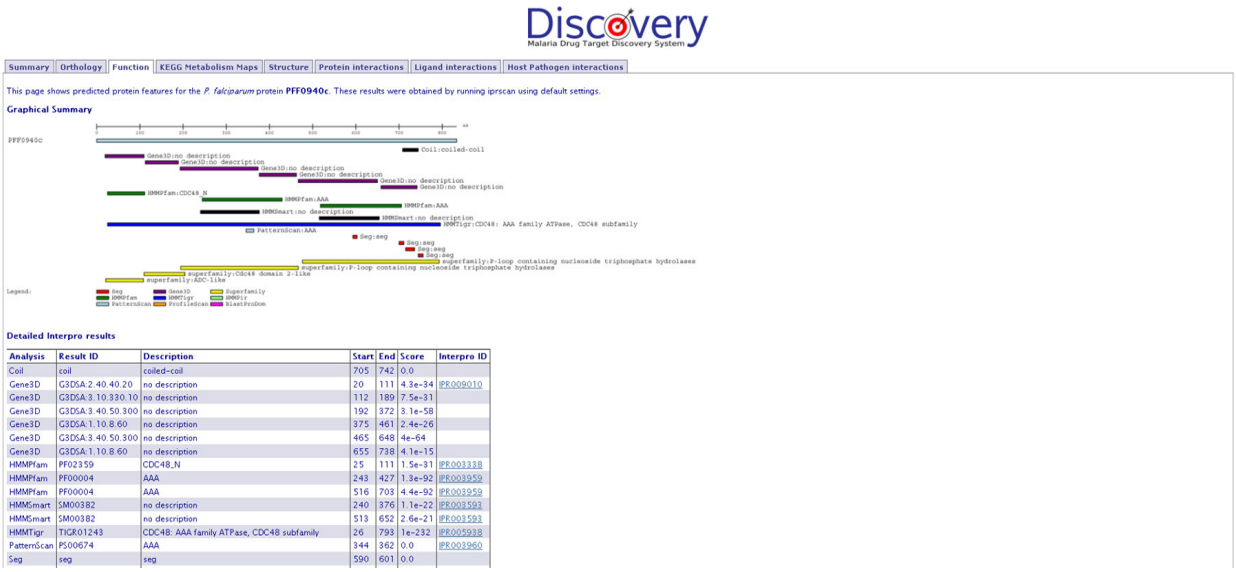

Figure 2.5 Function tab, containing results from INTERPRO SCAN

The KEGG **metabolism** maps tab utilizes EC nomenclature to identify the metabolic maps in which the queried protein takes part. In the case of this protein, no metabolic map was found (figure 2.6).

Figure 2.6 Metabolic map results of PFF0940c

For illustrative purposes, a search was repeated with a better annotated *Plasmodium* protein, PFD0830w. PFD0830w's metabolic map results are shown in figure 2.7.

As depicted in the results, PFD0830w is active during the folate biosynthesis pathways. DISCOVERY highlights the protein's specific enzymatic activity with a yellow block, which can be seen in the bottom right corner of figure 2.7. PFD0830w's enzyme nomenclature is 1.5.1.3, which means it falls within the *Oxidoreductases* category enzyme reactions.

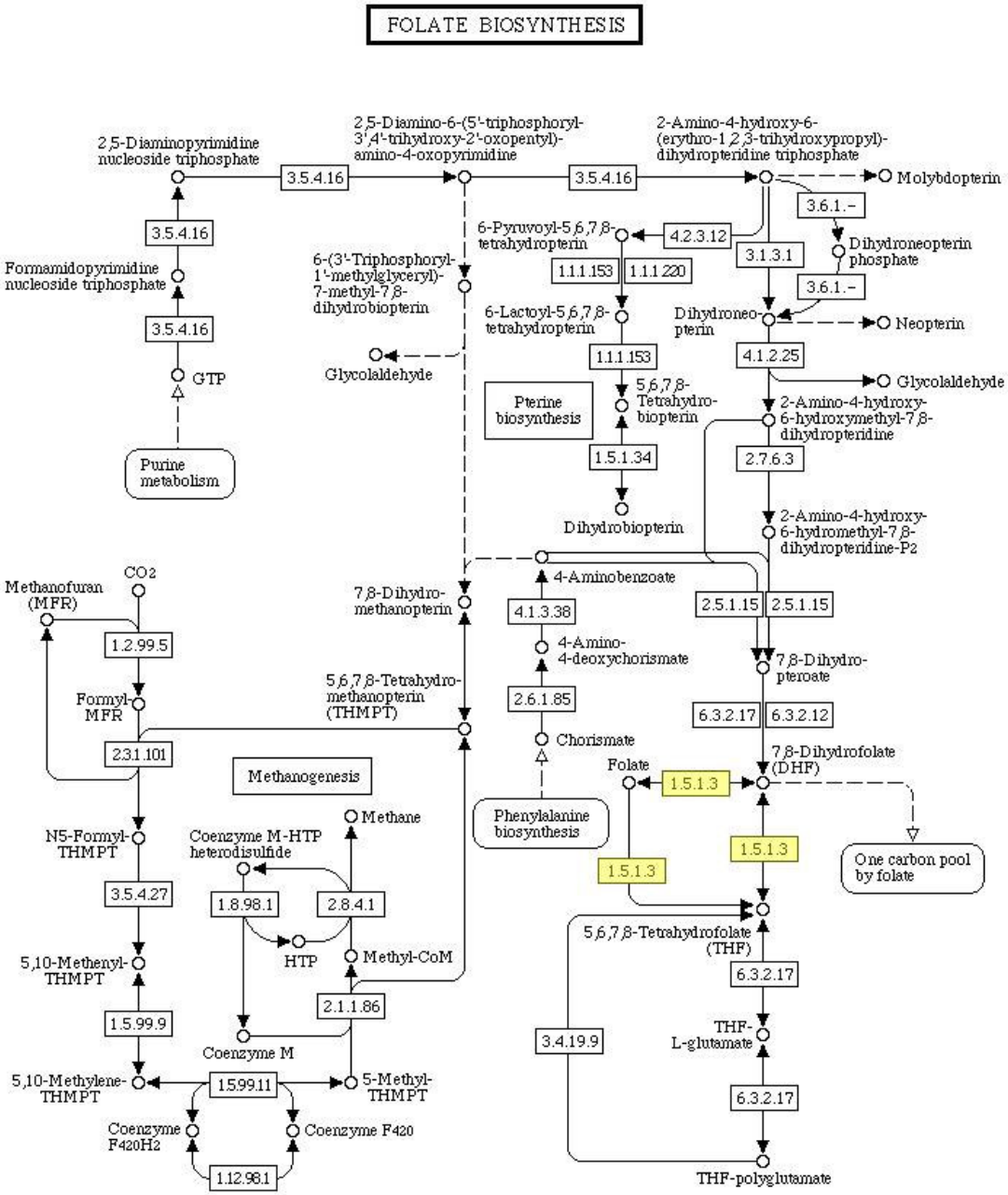

Figure 2.7 Metabolic map results of PFD0830w

Shifting the focus back to PFF0930c again, the structure tab contains results about **structural** analysis. Figure 2.8 shows two separate result tabs, MODBASE and PDB BLAST. The MODBASE tab reveals 4 structural models predicted for PFF0930c. These models are generated using MODBASE. The PDB BLAST tab contains a summary of BLAST vs PDB database results with their respective confidence scores.

Figure 2.8 Structure tab, containing structural predictions for MODBASE and BLAST vs PDB results

The **protein interactions** tab contains results from different sources (figure 2.9). The first source is experimental data from yeast two-hybrid studies from LaCount *et al.*, which predicts 3 experimental interactions with PFF0940c. One of these interactions predict that PFF0940c, cell division cycle protein, interacts with PF08_0003 a thryptophan / threonine rich antigen. This might mean that PF08_0003 is also important during cell cycle division. The next tab shows DISCOVERY'S ortholog-based predictions, using DIP and MINT protein-protein interaction databases. These predictions should be improved in future releases of DISCOVERY, since no additional filtering is applied during the prediction technique.

Figure 2.9 Protein interactions tab, containing protein-protein interaction information from *LaCount et al.* and DISCOVERY's own ortholog based predictions

Similar to the protein interactions tab, the **ligand interaction** tab contains information from experimental studies as well as predictions from DISCOVERY (figure 2.10). These interactions are sorted 'by ligand' and 'by source'. Several different ligand sources like KEGG, DRUGBANK, PDB, SMID, MSD were used during the identification of these interactions. The original ligand interaction predictions from SMID are expanded to include homolog-ligand interactions.

Figure 2.10 Ligand interaction tab, containing information from several ligand interaction sources, these interactions were expanded by indirectly linking interactions to protein homologs

The **host-pathogen interactions** tab contains *in silico* predictions using the protein-protein interaction database DIP and MINT, together with custom-made clusters generated via ORTHOMCL. Each of these host-pathogen interactions were evaluated according to,

- Sequence similarity
- Sub-cellular location
- PEXEL / VTS motif presence
- Microarray expression levels
- Metabolic map sharing

A score for each predicted interaction was calculated according to the above factors. This score depicts a measure of the likelihood of *in vivo* occurrence. Figure 2.11 shows the host-pathogen prediction results with their respective scores.

Figure 2.11 Host-pathogen predictions tab, containing DISCOVERY's ortholog based host-pathogen protein-protein interaction predictions

The architecture and implementation of predicting host-pathogen interactions will be discussed in detail in the following sections.

2.1.1.2 Keyword search

Another access method allows for mining of data using keyword searches. These keywords searches are based on descriptions of proteins, ligands, GO terms, EC numbers and interactions. The keyword search also contains help on using wild card characters in a search. Like the refined search and the chemical search, the keyword search returns a list of proteins containing the defined keyword. This list of proteins is grouped according to species where the different species are divided into different tabs (figure 2.12). The user can specify to view the results of one of the proteins within the list by clicking on a link; this link will lead to the same results as a quick search.

Figure 2.12 Keyword search results categorized according to species

2.1.1.3 Refined search

The refined search, allows the user to do a refined keyword search, where keywords can be used to define a specific field. The refined search also allows searches to include information from protein-protein and protein-ligand interactions and the searching of information on specified species only. As mentioned earlier the refined search returns a list of proteins from which a particular protein can be chosen and viewed in more detail.

2.1.1.4 Chemical search

The last access method provides chemical searches, which allows the user to search against uploaded or sketched chemical structures. This search was made possible via CHEMAXON JSEARCH. The results of such a search are shown below (Figure 2.13).

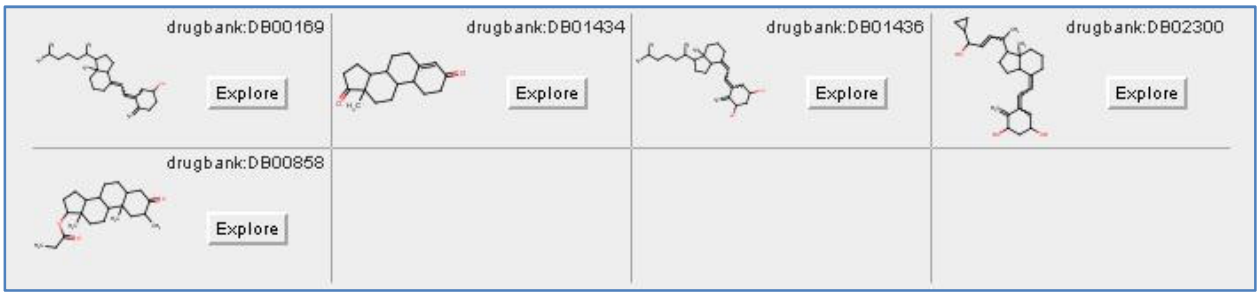

Figure 2.13 The results for finding compounds containing a pentameric ring, these results are limited to a maximum of 5 hits from DRUGBANK

By exploring a specific structure, DISCOVERY returns details predicting possible ligands as well as possible protein interactions.

2.1.2 TURBOGEARS

DISCOVERY was developed using TURBOGEARS (Ramm*, et al.*, 2007), a web development toolkit that consists of three basic components; the Controller, Model and View. Thus, the development for this study was also performed in the TURBOGEARS environment, using the Python programming language.

The Controller is the component responsible for the completion of all the tasks. The Controller also forms the part of the architecture that consists of a programming language, which allows the fulfillment of tasks. The Controller can communicate with both the Model and View components. DISCOVERY uses Python as programming language. The Model forms the backbone

of the network, which allows organization and communication with the database. The Model is used to create new tables and table-to-class mappers for handling data. Communication between the Controller and the Model allows the manipulation and viewing (via communication with the View component) of data. DISCOVERY makes use of a MYSQL (http://www.mysql.com) database for storage. Communication with the database is established with SQLALCHEMY (http://www.sqlalchemy.org/), which provides an object-based interface to the relational database. The view component allows the easy and effective communication between a user and a webpage.

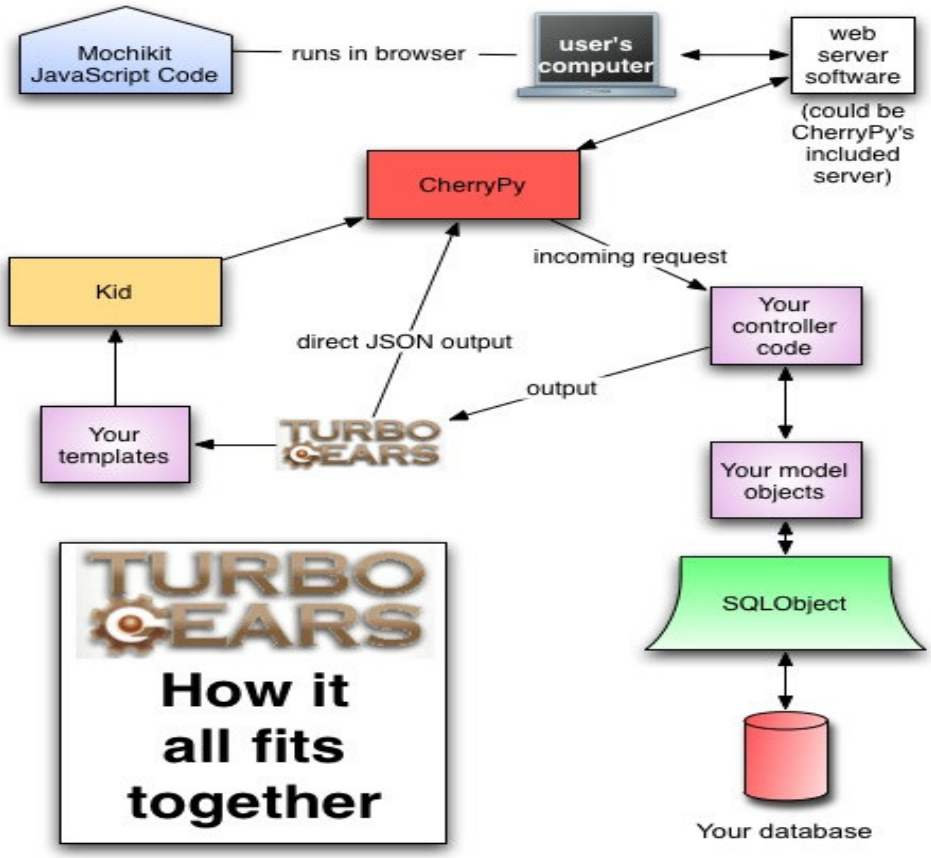

Figure 2.14 Flow diagram of the core units of TURBOGEARS
(http://www.swaroopch.com/files/200511/turbogears/howitallfits.jpg)

In a nutshell (Figure 2.14), user input is sent to the Controller *via* the View component; the Controller completes the tasks specified by the user. These tasks usually include storage or selection of data via the Model. The Controller then sends data or a message to the View component to communicate the status of the task to the user.

2.1.3 *In silico* prediction of host-pathogen interactions

Literature studies show only two other investigations that have aimed to predict host-pathogen interactions in malaria (Dyer *et al*. and Lee *et al*.). Similarly to Lee *et al*, DISCOVERY's host-pathogen predictions are based on an ortholog-based approach.

ORTHOMCL uses integration of information from a single organism with information from multiple different organisms *via* the relation of orthologous or in-paralog genes between them. Previous studies have proven that orthologs reveal considerable functional similarity. This approach has been implemented successfully in predicting protein-protein interaction networks like HOMOMINT (Persico*, et al.*, 2005) and ORPID (Brown and Jurisica, 2005).

Various methods of ortholog prediction exist. It is important to know the difference between orthologs and paralogs when deciding which ortholog prediction tool to use. Orthologs are defined as the evolutionary relationship between homologous genes that originated from a speciation event, while paralogs are homologous genes that originated from a common ancestor through gene duplication events (Fitch, 1970). Importantly, orthologs are more likely to be conserved and therefore to retain their functionality. While paralogs often mutate which lead to either new acquired functions or the loss of functionality, over evolutionary time. Considering this factor of change over time, it could be argued that recent paralogs or in-paralogs would not have undergone much mutation and therefore would also retain functionality. Thus, for functionality studies it makes sense to separate in-paralogs and

orthologs from out-paralogs or evolutionary older paralogs that would have undergone change (Li*, et al.*, 2005).

Another crucial factor is the comparison of the consistency, reliability and the biological significance of the different prediction methods. According to Chen *et al*. methods based on BLAST tend to be more sensitive, whereas methods based on trees tend to be more specific (Chen*, et al.*, 2007). Their results reveal two methods that had the best overall balance between sensitivity and specificity (both specificity and sensitivity above 80%); these methods were INPARANOID (Remm*, et al.*, 2001) and ORTHOMCL (Li*, et al.*, 2003). Hulsen *et al.* stressed the point that a method's sensitivity and specificity means nothing if the predictions made have little or no biological relevance (Hulsen*, et al.*, 2006). In their study they compared all the current most popular ortholog prediction methods with each other based on the assumption that functionally equivalent orthologs should behave similarly in functional genomics data (Sjolander, 2004). They analyzed functional conservation by comparing expression profiles, molecular functions and pairwise conservation of functional parameters. Pairwise conservation of functional parameters was analyzed by a measure of co-expression, neighbouring relations and protein-protein interactions between two species. These features were tested on two species pairs *H. sapiens – M. musculus* and *H. sapiens – C. elegans*. An overall score was calculated; both INPARANOID (Remm, *et al*., 2001) and ORTHOMCL (Li, *et al*., 2003) again had scored within the top three positions (Table 2.1).

Table 2.1 Overall comparison score of most familiar ortholog prediction methods (Hulsen, *et al.*, 2006)

H. sapiens - M. musculus	**Overall score**	**Rank**
Best bidirectional hit	5.42E+16	2
InParanoid	5.57E+16	**1**
euKariotic Orthologous groups	3.61E+11	6
OrthoMCL	5.10E+16	**3**
Z I Hundred	1.56E+15	4
Phylogenetic tree	7.46E+13	5
H. sapiens - C. elegans	**Overall score**	**Rank**
Best bidirectional hit	5.00E+10	6
InParanoid	7.90E+12	**1**
euKariotic Orthologous groups	7.58E+11	5
OrthoMCL	1.91E+12	**2**
Z I Hundred	1.22E+12	4
Phylogenetic tree	1.27E+12	3

As a result INPARANOID and ORTHOMCL were investigated more closely. INPARANOID is based on similarity scores. Initially, the intra- and interspecies pairwise similarity scores are calculated. These scores are used to select potential ortholog pairs (bitscore ≥ 50 and overlap ≥ 50%). Best directional hits are determined from the potential orthologs. If an interspecies pair scored higher than the intraspecies, the interspecies (in-paralog) was also included as part of the orthologs. Eventually the overlapping orthologs are stacked and a bootstrap-based confidence score is calculated for each ortholog group. Although INPARANOID is the best ortholog predictor according to analyses (Chen, *et al.*, 2007; Hulsen, *et al.*, 2006), it is based on the assumption that all pairwise comparisons occur between only two species (Li, *et al.*, 2003). This makes INPARANOID less suitable for DISCOVERY's host-pathogen interaction predictions because it needs to determine orthologs between multiple species.

ORTHOMCL utilizes a Markov clustering (MCL) algorithm to group orthologs with recent paralogs. The MCL algorithm is designed to find clusters in graph structures. The probabilities of random walks through a graph is calculated and formulated into stochastic matrices. These matrices

can be transformed using two operators transforming one set of probabilities into another. This transformation process is used to record the random walks on a graph using Markov matrices and other mathematical concepts (Van Dongen, 2000). Similar to INPARANOID, ORTHOMCL also uses best hit reciprocal BLASTP to identify orthologs and paralogs, with the biggest difference being the classification of in-paralogs. Paralogs with BLAST scores higher than respective ortholog pairs are determined to be in-paralogs. Using classifed in-paralogs and orthologs a graph is constructed where each node of the graph represents a protein and the edges represents the relationship between the proteins. Initially edges are weighted according to the BLAST scores ($-\log_{10}$), but secluded higher scores for in-paralogs can bias these relationships. Therefore the edge weights are normalized to reflect an average weight of all the orthologs between the relevant two species. The normalization step is followed by clustering of the graph into smaller groups according to protein relationships (edges). This study used ORTHOMCL to predict orthologs.

2.2 Data resources used for host-pathogen protein-protein predictions in DISCOVERY

The following section discusses the major sources of data used in this study.

2.2.1 Interaction databases

Many different interaction databases exist. DIP, the database of interacting proteins and MINT (Chatr-aryamontri*, et al.*, 2007), the molecular interaction database were employed in this study. A literature review revealed one article that also makes use of an ortholog-based approach. Lee et *al*. gathered their interaction data from POINT and POINTNET.

2.2.1.1 DIP

DIP (Xenarios*, et al.*, 2002) is a relational database that serves as storage space for experimental protein-protein interactions evidence. Various studies' results are combined into this single database. The manner in which DIP's data is stored, allows the identification of binary protein-protein interactions as well as multi-protein complexes. One concern is the comparability of the quality of experimental data from various experimental backgrounds; hence DIP utilizes quality assessment procedures. DIP is a semi-automatic curated database, with a combination of manual curation and computational data comparisons against a reliable CORE subset of DIP interaction data.

2.2.1.2 MINT

MINT (Chatr-aryamontri, *et al*., 2007) is a relational database as well. The MOLECULAR DATABASE OF INTERACTIONS contains a collection of functional interactions. It takes into account possible modification to one of the partners in an interaction. The interactions of MINT are experimentally verified information extracted from scientific literature by expert curators. The expert curators are assisted by the "MINT ASSISTANT" program, a text-mining program that focuses on identifying descriptive words concerning interactions or protein-protein relationships in the abstracts of scientific articles. After detection of each interaction, the interaction is scored based on a function of cumulative evidence that already exists in MINT.

2.2.2 General Protein resources

2.2.2.1 UNIPROT Knowledgebase

UNIPROT (Apweiler, *et al.,* 2004) acts as a central storage space for high quality protein sequence. UNIPROT consists of four sub-units, UNIPROTKB, UNIPARC, UNIREF and UNIMES.

The subset of interest for this study is UNIPROT Knowledgebase (UNIPROTKB). UNIPROTKB, is a non-redundant database that acts as the central integration point of all the protein data from the various sources.

UNIPROTKB consists of two sections, UNIPROTKB/SWISSPROT and UNIPROTKB/TREMBL. UNIPROTKB / SWISSPROT is based on manually curated literature annotation of proteins. Annotation is curated by experts to accomplish accuracy. A crucial step of annotation is the merging or stacking of different reports of the same protein. After careful inspection, a reference sequence is selected, any divergence from the reference sequences is then annotated and cross-references are provided. UNIPROTKB/TREMBL contains computationally determined annotations and classifications of proteins. To maintain high quality, automatic annotation uses rules on SPEARMINT, manual curation, HAMAP families, RULEBASE and PIRSF. UNIPROTKB/TREMBL consists of information gathered from translations of all coding sequences from databases like EMBL, GENBANK, ENSEMBL and TAIR.

2.2.2.2 UNISAVE

UNISAVE (Leinonen*, et al.*, 2006) forms part of the UNIPROTKB archive. As mentioned above UNIPROTKB contains data from both SWISPROT and TREMBL. It is possible that the data entries in SWISSPROT/TREMBL may change as protein annotations improve, in these cases the protein id gets updated as well. Only the newest ids are shown in SWISSPROT/TREMBL.

UNISAVE is a database that lists all the versions of the ids in SWISSPROT/TREMBL. This connection is typically needed for proteins referenced within older articles. UNISAVE is essential when comparing DISCOVERY's predicted host-pathogen protein-protein interactions with the previous predictions from Dyer *et al*. and Lee *et al*.

2.2.2.3 REFSEQ

REFSEQ (Pruitt*, et al.*, 2007) is an **almost** non-redundant collection of sequences representing genomic data, transcripts and proteins. The aim of REFSEQ is to be a reference for proteins from any given species. As of 2009, REFSEQ consist of about 2400 organisms, with a taxonomic reach from eukaryotes and prokaryotes to viruses. NCBI constructed REFSEQ, using data from GENBANK'S archives. REFSEQ sequence curation and quality control must meet three requirements; sequences must have precise nucleotide-to-protein sequence translation, sequences must be in valid ASN.1 format and species must support collaboration with official nomenclature groups.

2.2.2.4 TAXONOMY browser

The NCBI TAXONOMY BROWSER (Sayers*, et al.*, 2009) is generated from GENBANK data. TAXONOMY BROWSER indexes over 55 000 different species. TAXONOMY BROWSER can be used to display different ranges of useful information about GENBANK; from the display of the taxonomic position of a protein to the amount of sequences and structures that are available for specific species. For the purposes of the current study TAXONOMY BROWSER is only needed to enable collection of an organism name given a specific taxonomy id.

2.3 Analyses used for scoring of in vivo host-pathogen protein-protein predictions in DISCOVERY

Various methods to analyse proteins and protein-protein interactions exist. Computerized based characterization of proteins is a great aid in gathering knowledge on the regulation and functioning of proteins. DISCOVERY's *in vivo* predictions rely on integrated scores calculated from scores and weights of various analysis methods. These analysis methods are summarized below (Figure 2.15).

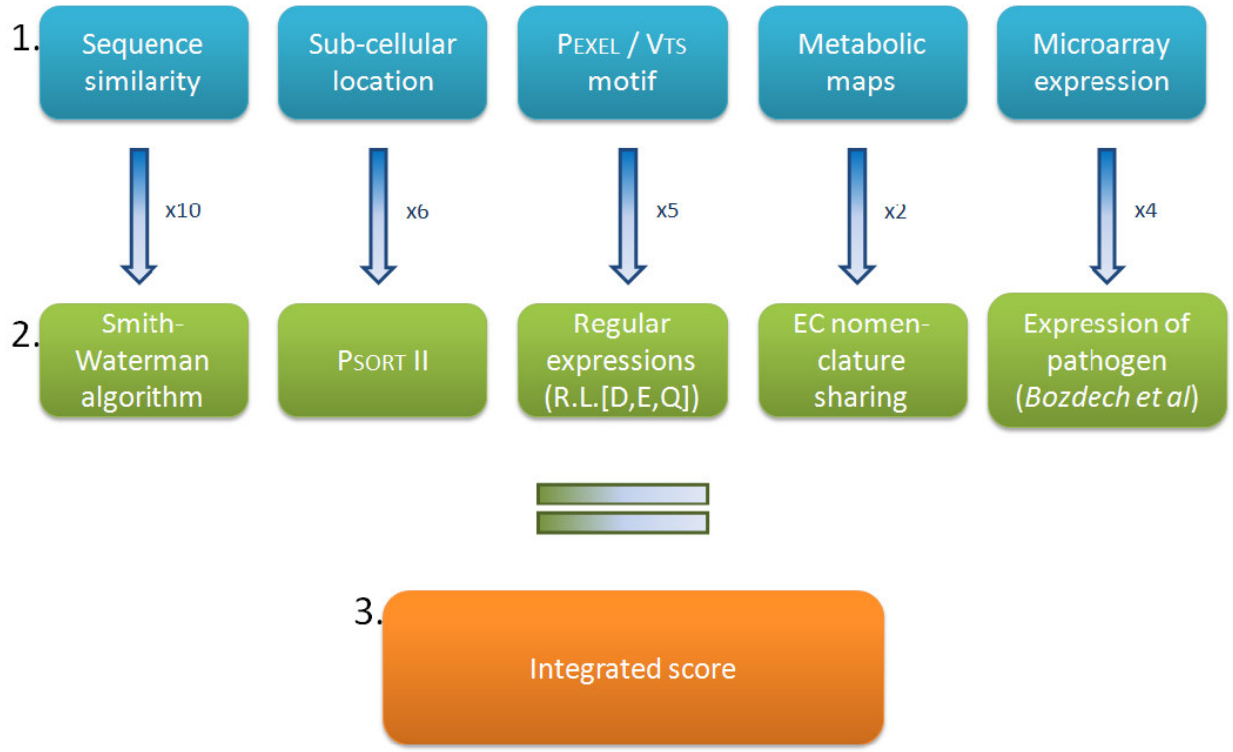

Figure 2.15 The analysis features used to determine *in vivo* interactions are listed in the first row, the arrows indicate the weight assigned to each of these features and the second row lists the methods that were used to calculate the integrated score for each interaction

2.3.1 Sequence similarity

The most popular method to predict the function of an unknown sequence is to determine the similarity of the sequence to annotated proteins or genes. This kind of search is called a homology search. Homology is recognized as two nearly similar sequences that share a common ancestor. The assumption on common ancestry is that the sequences also have a similar function.

The determination of sequence functionality needs to develop to faster and more accurate methods. Accomplishing this, would enable large scale genomic analyses. As usual the sensitivity and accuracy of a method plays an important role. In the case where search results contain all the sequences within a database as homologs, the method has complete sensitivity.

High sensitivity is usually accommodated by low accuracy; such methods lead to the misinterpretation of false positives as true positives. Very rigid methods on the other hand would exclude a lot of close similarities, but would contain only true positives.

Sphaer *et al.* (1997) conducted a study measuring the costs and benefits of sensitivity and accuracy of three well-known homology methods. These methods were BLAST (Altschul*, et al.*, 1990), FASTA (Pearson, 1990) and the dynamic programming-based method, Smith-Waterman (S-W) algorithm (Smith*, et al.*, 1985).

BLAST 2 and FASTA are known as heuristic methods that deduce possible good alignments, not taking all possible pairwise alignments into account. Heuristic methods' limitation in sensitivity ensures a huge gain in the speed of a search. The FASTA method **scans** the database for short identical matches and extends them using the S-W algorithm. BLAST2 is several times quicker than FASTA methods, although it is designed on similar principles. BLAST2 also starts-off with short identical sequences, but Karlin-Altschul statistics are used to extend these sequences. The S-W algorithm is somewhat slower than these methods, but it is also more sensitive. Due to its slowness, S-W is used less than the other methods. According to Shpaer *et al*. (1997) the use of faster methods provides less accurate results. Sphaer *et al.* (1997) claimed that BLAST and FASTA are less accurate than S-W because they exclude too much significant data.

2.3.1.1 Smith-Waterman similarity (S-W)

DISCOVERY utilizes S-W only to determine the similarity (Figure 2.16) between the orthologous proteins of the predicted host-pathogen protein-protein interactions and the orthologous proteins of interaction evidence from the global databases (DIP and MINT).

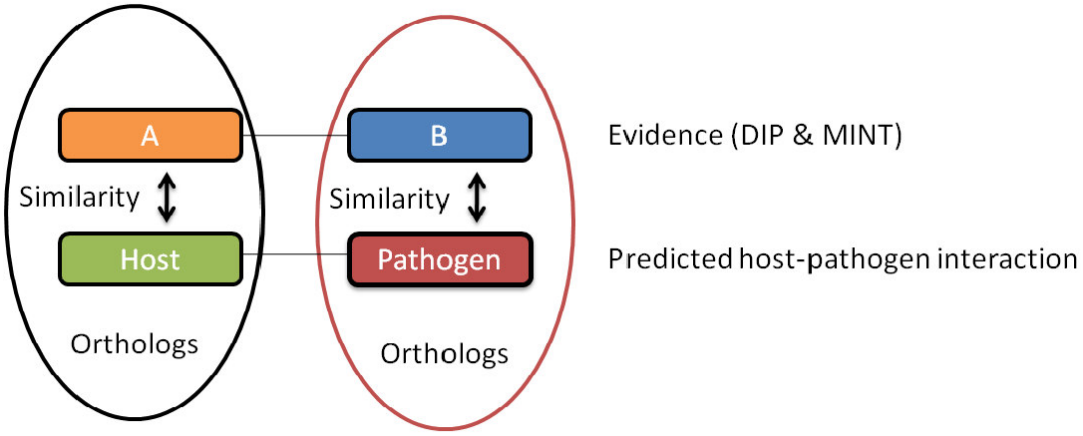

Figure 2.16 An illustration of how similarity is determined between the orthologous proteins of the predicted and evidence interactions

A higher similarity between these interactions would imply that the orthologous proteins within the interactions have similar functions and therefore would interact. The interactions are therefore weighted according to this similarity.

2.3.2 Sub-cellular location prediction

The environment in which a protein is found plays a pivotal role in a protein's secondary structure. Protein structure in turn is used to determine possible functioning of a protein, as well as possible interactions with other proteins (Drawid and Gerstein, 2000; Eisenhaber and Bork, 1998). DISCOVERY makes use of prediction tools to determine interacting proteins' sub-cellular locations. The predicted host-pathogen protein-protein interactions are then scored according to the probability of the interaction actually taking place *in vivo*. Assuming that interactions occur if the interacting proteins share a sub-cellular location, this score may be seen as a probability score that measures the likelihood that a host protein having a certain sub-cellular location would interact with a pathogen protein in the same sub-cellular location as the host protein.

Sub-cellular location prediction within prokaryotes is relatively simple as only 3 basic locations are predicted. These locations are allocated to the cytosol, integral membrane protein and exported proteins. In contrast with prokaryotes, eukaryotes have various sub-cellular location possibilities. A eukaryotic cell consists of numerous organelles, each having their own membranes. These different organelles can be seen in figure 2.17.

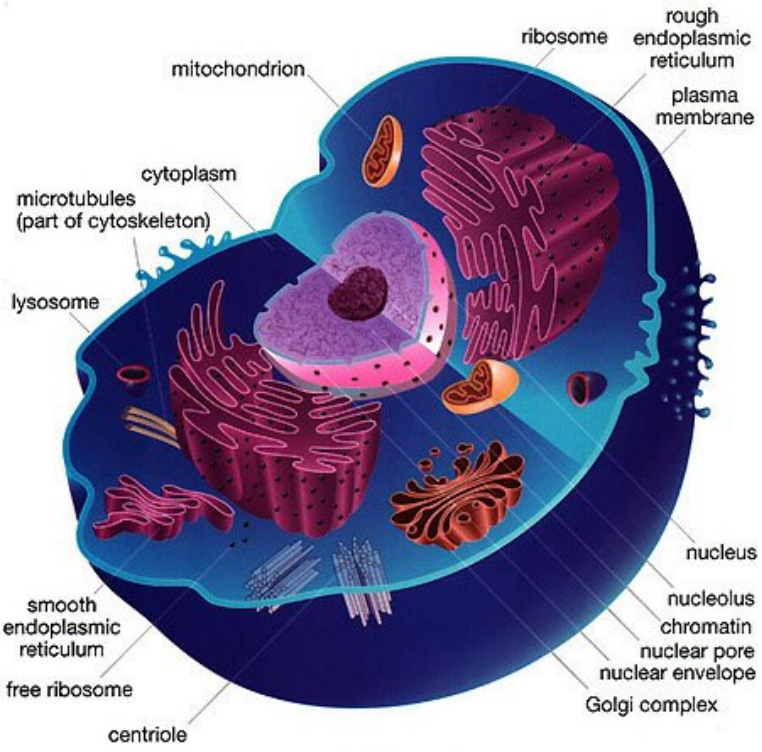

Figure 2.17 Organelles of an animal eukaryotic cell
(http://www.uvm.edu/~inquiryb/webquest/fa06/mvogenbe/Animal-Cell.jpg)

2.3.2.1 PSORT II

Various methods exist to predict the sub-cellular location of proteins. Most of these are based on either manually constructed explicit rules or rules modeled by automatic data-driven machine learning (SVN, ANN and HMM) algorithms. Three typical frameworks for prediction exist.

- Amino acid composition based prediction, where each amino acid's biochemical features are taken into account. Depending on the environment of a protein, protein folding occurs according to fixed motifs. These motifs are based on the principal of minimal energy.
- Expression profile based prediction, where proteins are grouped together because they contain a certain motif (Ex. peptide signals). These groups are characterized as belonging within a particular sub-cellular location.
- Phylogenetic profile based prediction, where shared ancestry between closely related proteins means the sharing of a sub-cellular location.

When deciding on a sub-cellular location prediction tool it is important to identify whether a method is a binary predictor or a multi-categorical predictor. Binary predictors aim to determine if a given protein belongs to a certain sub-cellular location or not. Binary methods tend to be highly sensitive, but not very specific which introduces the problem of predicting false positives. For this reason it was decided to rather use a multi-categorical predictor.

A multi-categorical predictor sorts a protein according to various sub-cellular locations at a time which makes it a more specific method. Accordingly, the more sub-cellular locations a method predicts the more specific it should become. According to an article that compares different sub-cellular location predictors with one another, PSORT is described as the gold standard of the field (Emanuelsson, 2002). PSORT makes predictions based on a set of sequence-derived constraints together with a few localization rules collected from literature. This includes localization of proteins according to sequence motifs. Thus, PSORT aims to integrate existing frameworks of sub-cellular prediction methods. DISCOVERY uses PSORT II to predict sub-cellular locations. PSORT II predicts for eleven different sub-cellular locations based on a k-nearest neighbour classification model. This makes PSORT II a sensitive and relatively specific method to use.

2.3.3 The PEXEL/VTS motif

The success of a pathogen is dependent on its ability to survive in a host. Pathogens have various mechanisms of evading a host's defence reactions. During the time that a pathogen is present in a host, the pathogen depends on the host to feed and survive.

In the case of malaria, *Plasmodium* survival depends on its ability to remodel the host erythrocyte (Boddey*, et al.*, 2009). An erythrocyte lacks metabolic processes, it is therefore necessary for a *Plasmodium* to utilize some kind of transport mechanism to feed on a host's resources. Hence, mechanisms to transport proteins within the host or to remodel the host cell to enable transport are crucial for P*lasmodium* survival. The *Plasmodium* depends on a parasitophorous vacuole (PV) for transport of proteins between itself and the erythrocyte. The development of the PV is illustrated in Figure 2.18. *Plasmodium* develops within the PV during the ring (0–24 hours), trophozoite (24–36 hours) and schizont stages (40–48 hours).

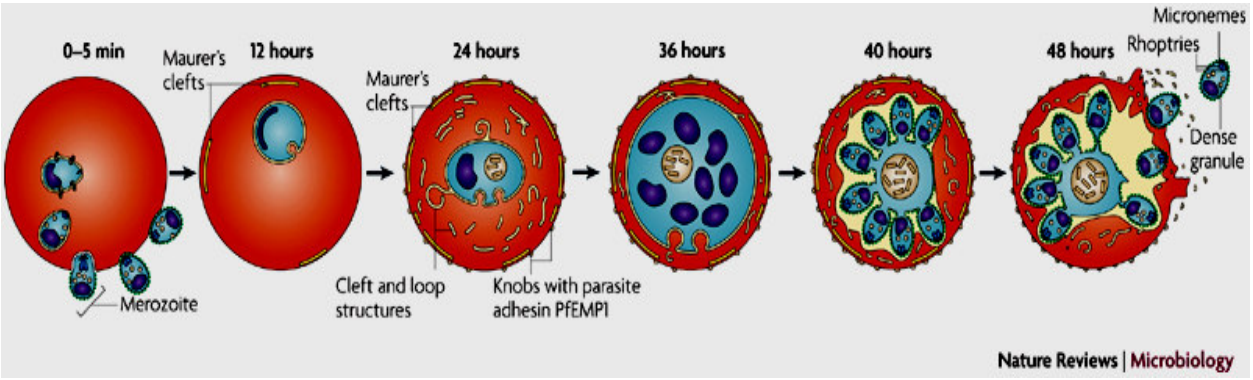

Figure 2.18 The different stages of *Plasmodium* parasitophorous vacuole development
(http://www.nature.com/nrmicro/journal/v7/n5/images/nrmicro2110-f1.jpg)

During the fully developed stages (24 hours) membrane-bound structures are formed within the cytoplasm of the erythrocyte. These structures develop into daughter merozoites which are later released into the haemoglobin when the erythrocyte cell ruptures (16–32 hours). The

PV not only protects the *Plasmodium* parasite during development, but also allows for the transport of proteins into the erythrocyte. Studies have determined that proteins that get transported through the PV contain a pentameric motif (R.L.[D,E,Q]), that is called the vacuolar transport signal or the *plasmodium* export element (VTS / PEXEL) (Horrocks and Muhia, 2005).

Proteins that are secreted by a pathogen are some of the most important proteins studied during drug manufacturing, because these proteins have a greater likelihood to interact with the host proteins.

It is therefore of great importance to determine the presence of PEXEL / VTS motif within each of the *Plasmodium* proteins in the DISCOVERY system. Hence, a predicted host-pathogen protein-protein interaction that contains a PEXEL / VTS motif has to be promoted above predictions without the presence of a PEXEL / VTS motif.

2.3.4 Metabolic maps with the Enzyme Commission (EC) classification

The EC commission (EnzymeNomenclature) contains numerical classifications of enzyme interactions. The EC classifies interactions into six categories (Table 2.2). Each of these categories is sub-categorized into smaller groups. These EC categories are categorized over different species, meaning that if two proteins have the same EC number their origin are disregarded.

Table 2.2 Main categories of the EC classification (http://expasy.org/enzyme/)

Category	Type of reaction
EC 1	*Oxidoreductases*
EC 2	*Transferases*
EC 3	*Hydrolases*
EC 4	*Lyases*
EC 5	*Isomerases*
EC 6	*Ligases*

Figure 2.19 shows an example of a metabolic pathway.

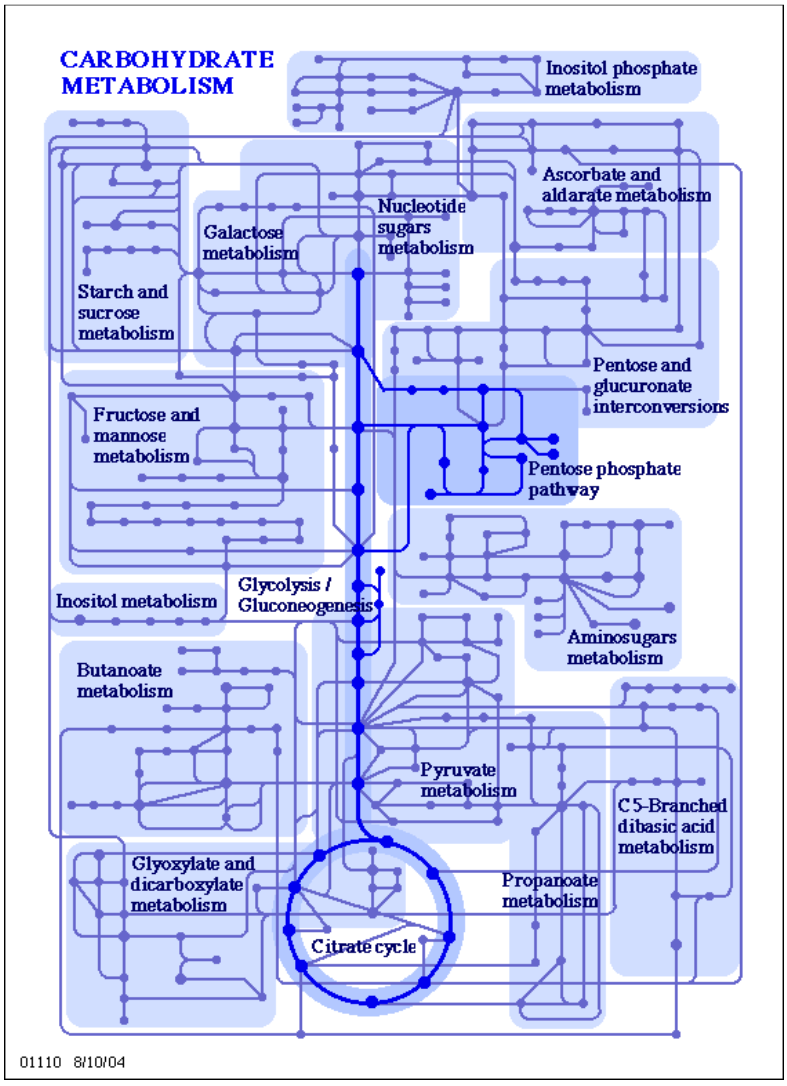

Figure 2.19 A metabolic pathway of carbohydrate metabolism
(http://media.wiley.com/CurrentProtocols/BI/bi0112/bi0112-fig-0005-1-full.jpg)

A metabolic pathway usually consists of various metabolic maps, for example in Figure 2.19 each node within the metabolic map of the citrate cycle represents a protein-protein interaction. Each node contains an EC (Enzyme commission) number. Consequently, proteins that share a metabolic map are possibly more likely to interact than proteins that do not.

Thus, if a predicted host-pathogen protein-protein interaction contains proteins that share an EC number, this prediction is potentially more likely to be a true positive and should therefore be preferred above an interaction where this isn't the case. Sharing of an EC number could typically be described by a presence of bifunctional enzymes. Bifunctional enzymes usually contain two structural subunits. Meaning that enzyme activation requires the integration of the two subunits, which allows the enzyme substrate to be regulated in metabolic pathways (Moore, 2004). In turn enzyme products can directly or indirectly regulate enzyme activation; substrate feedback allows inhibition of enzyme activity by means of substrate challenging. The identification of such substrates may prove useful for drug design and discovery.

In DISCOVERY EC numbers for all the host and pathogen proteins were determined using the EC commission system. EC numbers are used in attempt to exploit possible substrate challenging interactions between host and pathogen proteins.

2.3.5 Experimental microarray results

Microarrays are very useful for determining a protein's regulation and biochemical function. This kind of information is crucial during the prediction of possible drug-hits and the development of vaccines.

Bozdech *et al*. (2003) analyzed the complete asexual intraerythrocytic development cycle (IDC). The processes during the IDC are extremely important for the survival of *Plasmodium.* Accordingly, knowledge on the regulation of the proteins involved in these processes is fundamental for drug discovery. In their analyses Bozdech *et al.* uses the familiar chloroquine-sensitive HB3 strain (Walliker*, et al.*, 1987; Wellems*, et al.*, 1990) of *P. falciparum* and hybridizes it with a reference strain 3D7 which contains all the developmental stages of *Plasmodium*.

The analyses suggest that 60 % of the *Plasmodium* genome is transcribed during the IDC. After multiple sorbitol treatments the HB3 strain was synchronized and a time-map for each phase in the IDC could be established. It took the *Plasmodium* parasite approximately 2 hours to infect the erythrocytes. Infection is followed by a ring phase, trophozoite phase and a schizont phase which respectively occurs at 0 -17 hours, 17-29 hours and 29-53 hours (Figure 2.20).

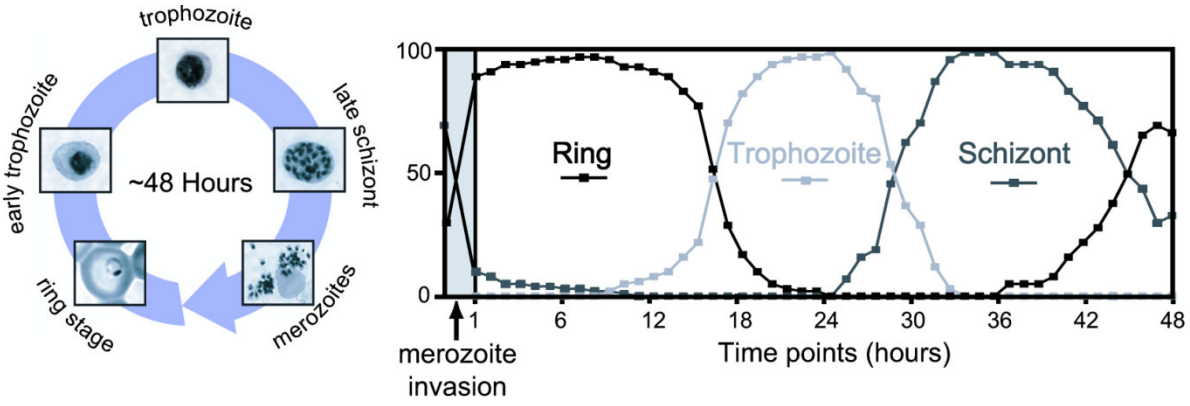

Figure 2.20 Timeline for the asexual intraerythrocytic development cycle
(Bozdech, *et al.*, 2003)

Evidently Bozdech *et al.* have identified 300 transcriptionally active genes in the final hours of the IDC that corresponds to transcribed genes in the early-ring phase. This implies the continuous cycling of the IDC.

The graph surface of a pathogen protein expression levels within the IDC indicates a protein's activity. Activity of proteins are regulated by protein interactions, often host proteins inhibit or enhance pathogen proteins. If a pathogen protein that takes part in a predicted interaction has high activity during the IDC, the likelihood that this protein may interact with a host protein increases. This way DISCOVERY uses expression data to add to the discriminating power of the predicted host-pathogen protein-protein interactions.

2.3.6 The NEGATOME

The NEGATOME is a database that consists of interactions that are unlikely to occur physically. The database currently contains interactions that are proven not to exist according to experimental results. Information about these non-interacting proteins is mined *via* manual curation of literature searches and analyses of proteins' three dimensional structure.

The NEGATOME is used in DISCOVERY to exclude host-pathogen protein-protein interactions that are incorrectly predicted. Although the NEGTOME currently only contains a limited amount of data and therefore would not remove all the false positive interactions, it remains a useful analyses method for measuring the accuracy of DISCOVERY's ortholog-based host-pathogen protein-protein interaction predictions.

2.4 Data integration and implementation of the host-parasite interaction prediction methods

The following section provides technical detail regarding the integration of the various relevant data types, and the implementation of the host-parasite protein-protein interaction prediction methodology. The prediction process can be subdivided into the three main steps depicted in Figure 2.21.

The first step involves generating custom made clusters from annotated data already existing in DISCOVERY and data from DIP and MINT, via ORTHOMCL. This is followed by the integration of species data, cluster data, annotated data from DISCOVERY and DIP and MINT to predict possible *in vitro* host-pathogen interactions. The final step involves the scoring of predicted interactions using to sequence similarity, sub-cellular location, microarray expression, PEXEL / VTS presence and metabolic pathway sharing to determine the likelihood of the interactions occurring *in vivo*.

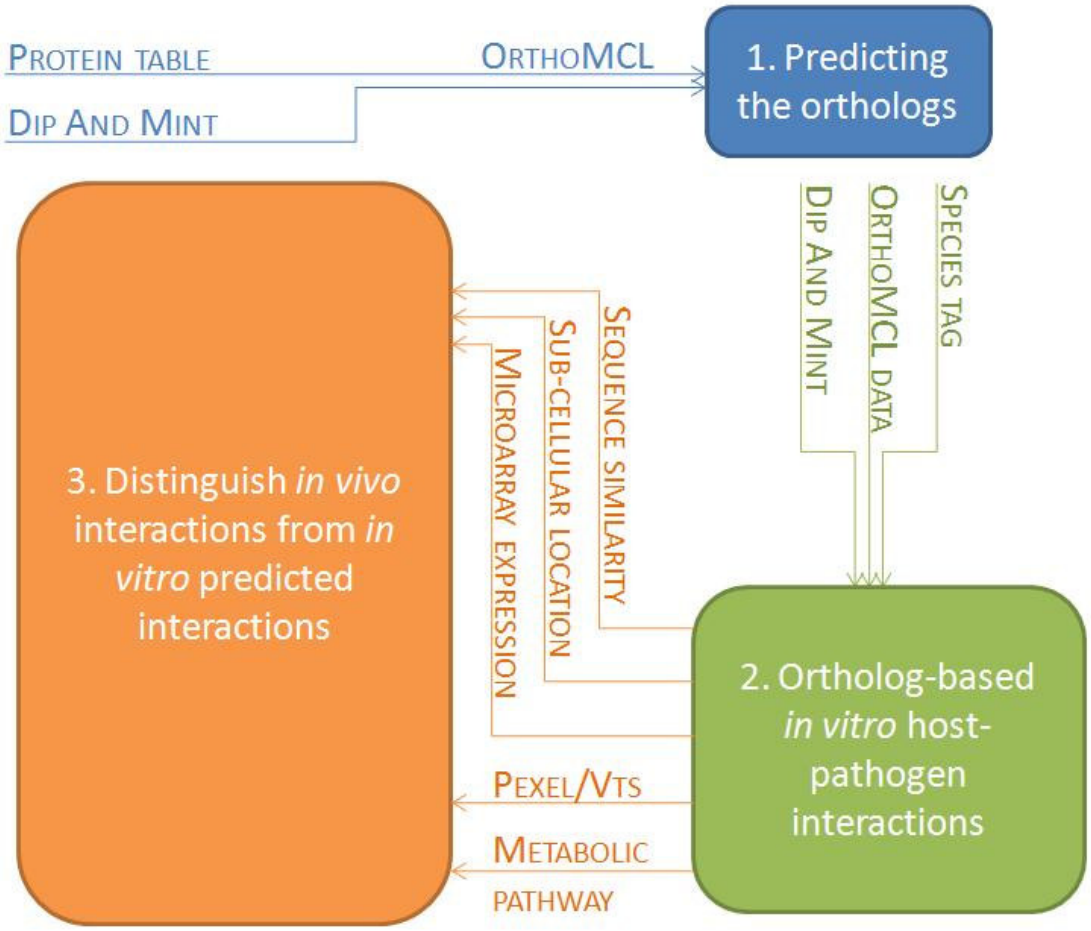

Figure 2.21 The mains steps followed in predicting host-pathogen protein-protein interactions in DISCOVERY

2.4.1 Predicting the orthologs

The public interaction databases DIP and MINT are parsed and stored in the *interactionsummary* table (Figure 2.22). The organism data in this table has been mined from TAXONOMY BOWSER using the taxon ids from DIP and MINT as reference. Importantly no sequence information for the proteins in the interaction databases was available in the interaction databases themselves. As discussed later, these sequences were retrieved from UNIPROT and REFSEQ and stored in a separate table.

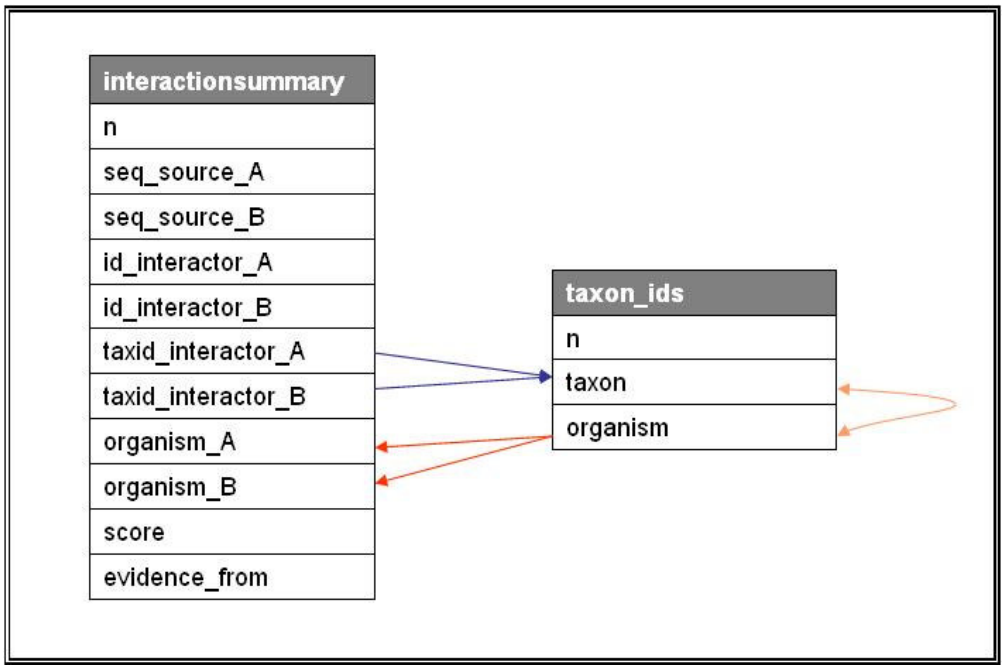

Figure 2.22 The 'interactionsummary' table consists of a combination of DIP and MINT data taxon ids and organism details were gathered from TAXONOMY browser.

If an interaction within DIP or MINT contained multiple ids per interactor, as seen in Figure 2.23, it is called a complicated interaction. These interactions are simplified into multiple unique interactions.

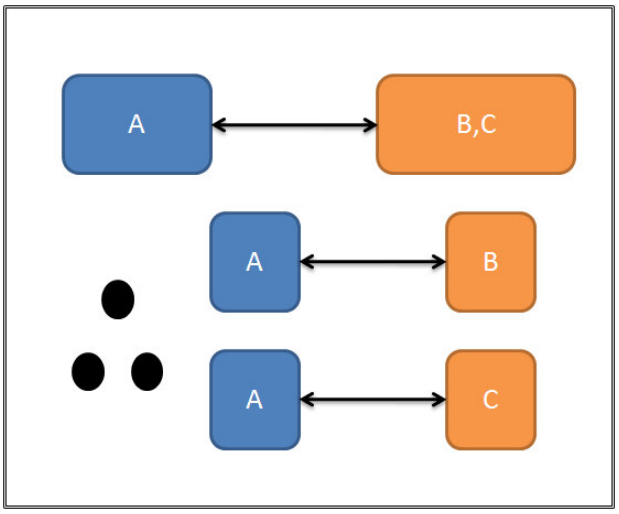

Figure 2.23 Simplification of interactions where one protein (protein A) interacts with multiple other proteins (prtoein B and C) into single unique database records with a basic 1:1 relation of proteins rather than 1:many

After extracting and storing all the experimentally proven interactions into the *interactionsummary* table each interactor and its sequence is stored into a new table called, the *interactors* table. The sequences for DIP and MINT were retrieved from UNIPROT and REFSEQ using the *id_interactor* column (UNIPROT or REFSEQ accessions) from the *interactionsummary* table (Figure 2.24). UNIPROTKB is stored into two tables called *swiss_prot_ids* and *trembl_ids* and the REFSEQ data is stored within the *refseq_ids* table.

The *interactors* table contains unique entries only, in other words it excludes all the interactor duplicates from the interactions in the *interactsionsummary* table.

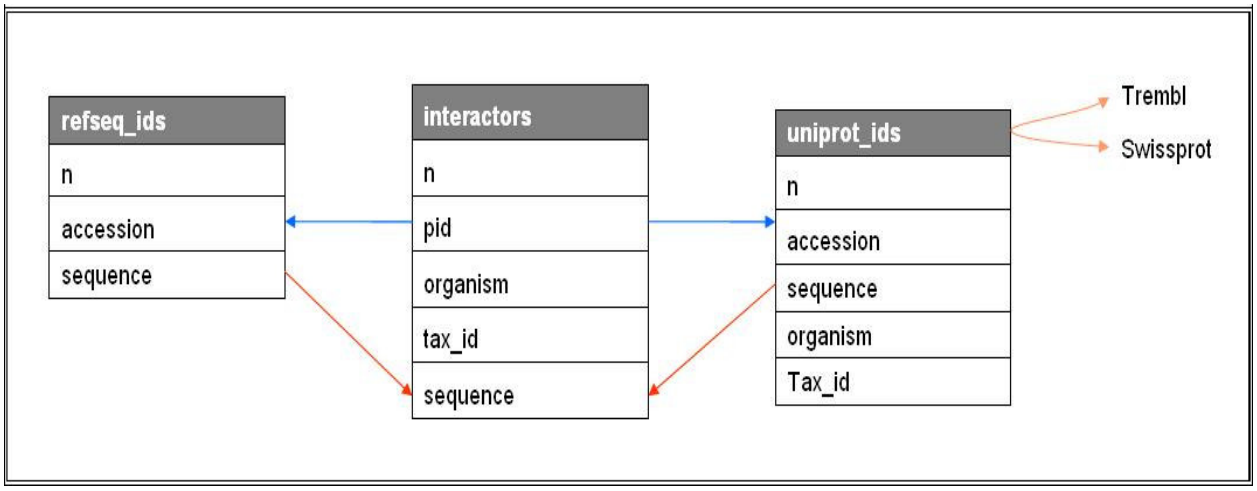

Figure 2.24 The 'interactors' table utilizes unique accessions (pid) to gather sequence information from UNIPROT Knowledgebase and REFSEQ.

Now data from *interactors* table and data from a prior existing table, the *protein* table could be combined into a single *fasta* file (Figure 2.25).

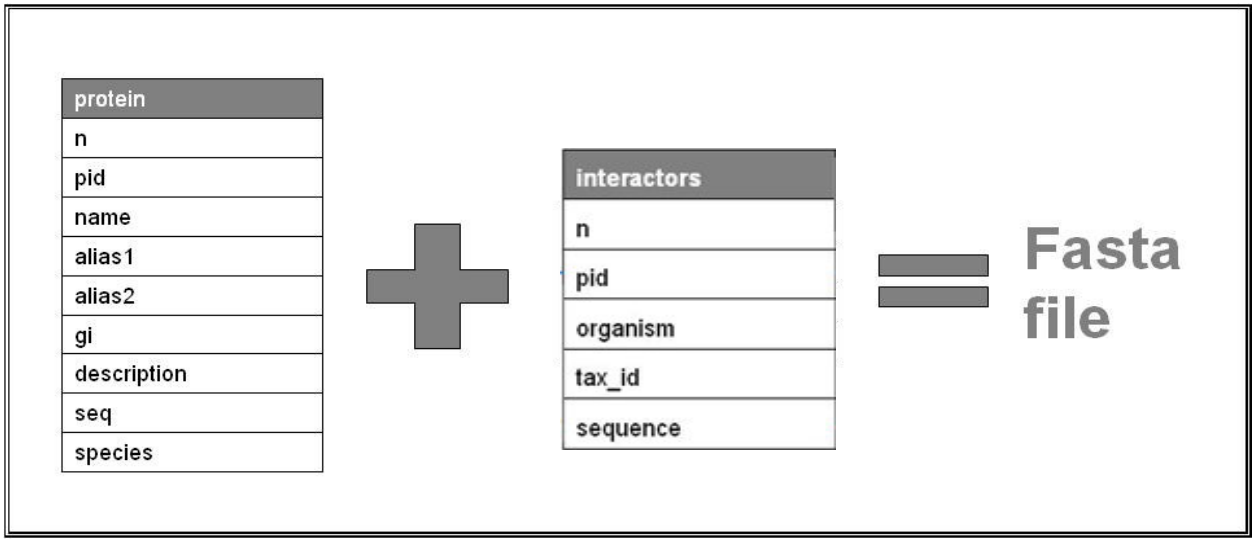

Figure 2.25 A fasta file is constructed using the combined data from the 'protein' table and the 'interactors' table

This *fasta* file is used to do ortholog clustering with ORTHOMCL. ORTHOMCL was run in mode 3 with the default parameters. Mode 3 takes fasta files as input and returns a file that contains protein accession numbers sorted into clusters. These clusters are ordered in descending order according to cluster size; by default cluster 0 contains all the protein accessions. All the clusters except for cluster 0 are stored in the *orthomcl_interactions* table (Figure 2.26).

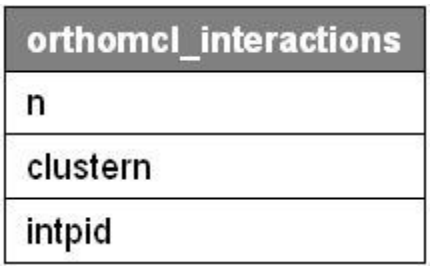

Figure 2.26 The OrthoMCL results table contains only a cluster number and a protein accession

The integration of the ORTHOMCL results with the host and pathogen data from the *protein* table in DISCOVERY leads to DISCOVERY's host-pathogen protein-protein interactions predictions. The exact technique will now be explained.

2.4.2 Ortholog based *in vitro* host-pathogen interaction prediction

In DISCOVERY the ortholog based host-pathogen interaction predictions are triggered by a user-defined query. This query initiates a whole set of data queries from various tables in the DISCOVERY database. The whole process needs a protein ID as an initiator. The moment a protein is **queried**[1], the first step would be to check if the protein id **exists**[2] in the *protein* table. If the protein does exist in the *protein* table, the specified protein can either be a pathogen (*Plasmodium)* protein or a host (*H. sapiens* or *A. gambiae*) protein. A discussion on the prediction process will now follow by referring to Figure 2.27.

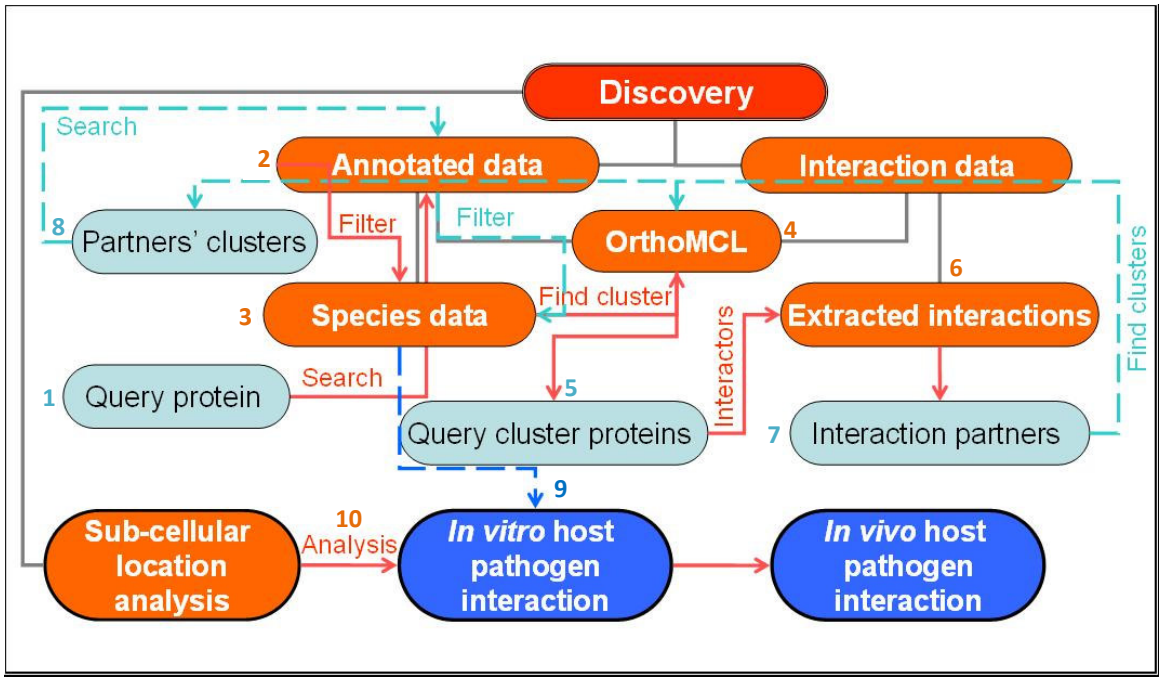

Figure 2.27 Methods involved in ortholog based host-pathogen interaction prediction

After a protein has been queried and checked for existence, the process will continue by **checking the species**[3] the protein belongs too. If a protein belongs to any of the *Plasmodium* species (added to DISCOVERY) the protein will be tagged as a pathogen, but if the protein belongs to *H. sapiens* or *A. gambiae* the protein is tagged as a host protein. The rest of the procedure depends on this tag.

2.4.2.1 Tagged as pathogen

After the protein is tagged as a pathogen, ORTHOMCL results are used to **determine** the protein's **orthologs**[4]. Subsequently each of these orthologs is used to **search** against the **public interaction databases**[5] (DIP and MINT). A search against such a large amount of interactions led to the reconsideration of the search procedure. Consequently, the interactions had to be filtered; after selecting interactions from MINT that has scores equal to and above 0.3 a further filter was implemented which only included interactions that existed in both the DIP and MINT databases[6].

If an ortholog matched one of the interacting proteins in the filtered interactions the focus was shifted towards the **other protein taking part in the interaction**[7]. Consecutively, each of these **interacting proteins' orthologs**[8] is determined. If these orthologs exist in the *protein* table, it once again needs to be classified either as a host or a pathogen protein. The host proteins are selected.

Due to this ortholog-based approach the initial query protein is **predicted to interact**[9] with each of the identified host proteins. Accordingly the orthologous interactions from DIP and MINT serve as evidence for each prediction.

2.4.2.2 Tagged as host

When the initial protein is tagged as a host protein, the procedures that follow are quite similar to the case where the initial protein is identified as a pathogen, the only difference being the filtering of the species at the end. While the initial protein is tagged as a host the final selected interacting partners should be pathogen proteins.

This prediction illustrates the power of integration and filtering of data from various sources. It is important to note that the predicted interactions are *in vitro* interactions and that it can only be accepted as valid ***in vivo* interactions after further analyses**[10].

2.4.3 Distinguishing possible *in vivo* interactions from *in vitro* predicted interactions

Previous methods made use of filtering to distinguish *in vivo* host-pathogen interactions from predicted *in vitro* interactions (Dyer, *et al*., 2007; Lee, *et al*., 2008). DISCOVERY makes use of scores to identify *in vitro* host-pathogen interactions that are likely to occur *in vivo*. Assignment of scores rather than filtering out likely interactions makes DISCOVERY more sensitive and less likely to exclude of misinterpreted significant data. In contrast with other methods, DISCOVERY's scoring method utilizes a combination of various sources to proof possible *in vivo* occurrence of interactions.

The different sources / characteristics used are sequence identity, PEXEL / VTS motif presence, sub-cellular location, metabolic pathways and microarray expression levels. Each characteristic is assigned an initial arbitrary weight (Table 2.3). These arbitrary weights are based on the theoretical understanding of parasite biology, and would certainly require optimization during subsequent follow-up studies.

Table 2.3 Weights assigned to each protein characteristic according to relevance for Discovery's host-pathogen protein-protein interaction predictions

Characteristic	Weight
Sequence identity	10
PEXEL/VTS motif presence	5
Sub-cellular location	6
Metabolic pathways	2
Microarray expression levels	4

According to these weights an interaction can have a maximum weight of 27, because all the characteristic scores vary with values between zero and one. The characteristic scores are calculated for every possible interaction prediction.

2.4.3.1 Sequence identity

Sequence identity is a measure of similarity between sequences. This similarity was calculated using the Smith-Waterman algorithm. The higher the similarity between sequences, the greater is the likelihood that two sequences share the same functionality and interactions. Sequence similarity was determined between each host protein and its orthologs from the evidence (DIP and MINT), likewise the similarity between the pathogen protein and its orthologs was determined. Subsequently, the averages for each consecutive two similarity scores were calculated as shown in equation 2.1.

Equation 2.1 Calculation of similarity of a predicted interaction

$$Interaction\_similarity = sim(host, host\_ortho_i) + sim(pathogen, pathogen\_ortho_i)/2,$$
$$for(i_2, i_3, i_4, ..., i_n)$$

The score from the Smith-Waterman algorithm ranges from zero and one, where one means complete similarity.

2.4.3.2 PEXEL/VTS motif presence

The evaluation of the PEXEL/VTS motifs is a binary evaluation method, where a score of one means that a PEXEL/VTS motif is present in an interaction. The presence of a PEXEL/ VTS motif implies that a pathogen protein is transported through the *Plasmodium* parasite's PV membrane. Proteins that transport through this membrane are more likely to interact with host proteins. Scoring for PEXEL/VTS motif presence is relatively easy, since regular expressions could be used to find PEXEL/VTS motifs within the pathogen proteins.

If a pathogen protein contains this motif a PEXEL/VTS motif score of one is allocated to the interaction, if no PEXEL/VTS motif is found the score is zero.

2.4.3.3 Sub-cellular location sharing

Using sub-cellular locations to filter for *in vivo* protein-protein interactions in a single species are quite straight-forward. When two proteins are predicted to interact and also share a sub-cellular location, no physical constraints suggests that this interaction will not occur *in vivo*. There are exclusions from this rule; proteins occurring on membrane surfaces and exported proteins. Membrane proteins are highly interactive because they reside within cell cytoplasm and are responsible for transport through membranes. Exported proteins are also very likely to interact with proteins in most sub-cellular locations.

However, establishing the same principals of sub-cellular locations between different species is not as simple. It is not always feasible to predict that proteins which belong to different species and share the same sub-cellular will occur *in vivo*. Consider the obvious example of proteins sharing a nucleus as sub-cellular location; what would the likelihood of these proteins to interact be, if they belong to different species? DISCOVERY attempts to take advantage of the knowledge of sub-cellular locations by scoring interactions according to the likelihood of

proteins interacting between species. This novel approach can be quite a tedious task when little interaction data could be acquired between the species of interest. In cases like these the logical way to assign scores, would be to consider the amount of physical barriers these proteins need to overcome to actually interact. DISCOVERY uses a scoring table based on the likelihood of proteins crossing physical barriers / membranes which enables *in vivo* interaction (Table 2.4).

To predict sub-cellular locations DISCOVERY utilizes PSORT II. PSORT II distinguishes between eleven different sub-cellular locations. Table 2.4 show the scores assigned to PSORT II's predicted sub-cellular locations.

Table 2.4 Scores allocated to the different sub-cellular locations in Discovery

Sub-cellular location	Score
Extracellular including cell wall	1
Plasma membrane	1
Vesicle of secretory system	1
Vacuolar	0.7
Peroxisomal	0.7
Cytoskeletal	0.5
Cytoplasmic	0.5
Golgi	0.3
Endoplasmic reticulum	0.3
Mitochondrial	0.2
Nuclear	0.2

Using these assigned scores a score for each predicted interaction could be calculated as illustrated in equation 2.2.

Equation 2.2 Calculation of interaction sub-cellular score

$$Interaction\_subcell = (host_{subcell} + pathogen_{subcell})/2$$

The sub-cellular location score is calculated for each predicted host-pathogen protein-protein interaction.

2.4.3.4 Metabolic pathways

If proteins that are predicted to interact, share a metabolic pathway the likelihood that they truly interact should increase (Kalyanaraman and Jacobson, 2010). Scoring interactions according to the metabolic pathways is a binary method. DISCOVERY utilizes EC nomenclature to determine if the proteins taking part in an interaction shares a metabolic pathway. According to EC nomenclature proteins that has the exact same EC number, exist in the same metabolic pathway. A metabolic pathway score can either be zero or one, depending on the sharing of a metabolic pathway between the proteins that are predicted to interact. A score of one entails that interacting proteins share a metabolic pathway.

2.4.3.5 Microarray expression levels

Expression levels of proteins provide information about the activity of proteins during certain treatments or periods. DISCOVERY use microarray data from Bozdech *et al,* this analysis measures the protein expression levels of *Plasmodium falciparum* during the infection of the erythrocytes throughout the asexual IDC (Bozdech, et al., 2003). The relevancy of a predicted interaction based on these results, revolves around the calculation of the surface under the expression level graph of a *Plasmodium* protein in a predicted interaction. To calculate an overall expression level of each interaction, the surface under the graph for each phase (ring, throphozoite, schizont) in the IDC is calculated relative to the maximum expression per phase (surface). This calculation is seen below (equation 2.3).

Equation 2.3 Calculation of the surface of the *plasmodium* protein in a predicted interaction

$$microarray\_score = \left(\frac{plasmodium_{ring\_surf}}{\max(ring\_surf)} + \frac{plasmodium_{trophozoite\_surf}}{\max(trophozoite\_surf)} + \frac{plasmodium_{schizont\_surf}}{\max(schizont\_surf)}\right)/3$$

The assumption around the expression is that a protein that is up regulated more often will have a greater surface under an expression graph. Therefore, such a protein is more likely to be interacting with other proteins; if this particular protein is predicted to interact with a host protein, the interaction's likelihood slightly increases.

2.4.4 Complete host-pathogen protein-protein interaction score

After all these characteristic scores are calculated they are integrated with the weights from table 2.3 to assign the total score to each interaction according to equation 2.4.

Equation 2.4 Calculation of the total score for each predicted interaction

$$Totalscore = w_{sequence\_id}(score_{sequence\_id}) + w_{Pexel/VTS}(score_{Pexel/VTS}) + w_{sub-cellular}(score_{sub-cellular})$$

$$+ w_{metabolic}(score_{metabolic}) + w_{\boldsymbol{expression}}(score_{\boldsymbol{expression}})$$

After calculating each score the scores were plotted onto a histogram (Figure 2.28). The histogram had a skewed-right distribution, with a maximum score of 20.7 and a minimum score of 5.8. According to this histogram a reasonable amount of interactions has been highlighted as significant, but the true test of the significance of the predicted interactions would be to see if how well data from *Dyer et al.* and *Lee et al.* agrees with the predictions made in DISCOVERY. To actually validate these *in silico* methods, it must be confirmed against real empirical data. An investigation into the results obtained from these predictions, and a comparison against the results from Dyer *et al*. and Lee *et al*. is provided in the following chapter.

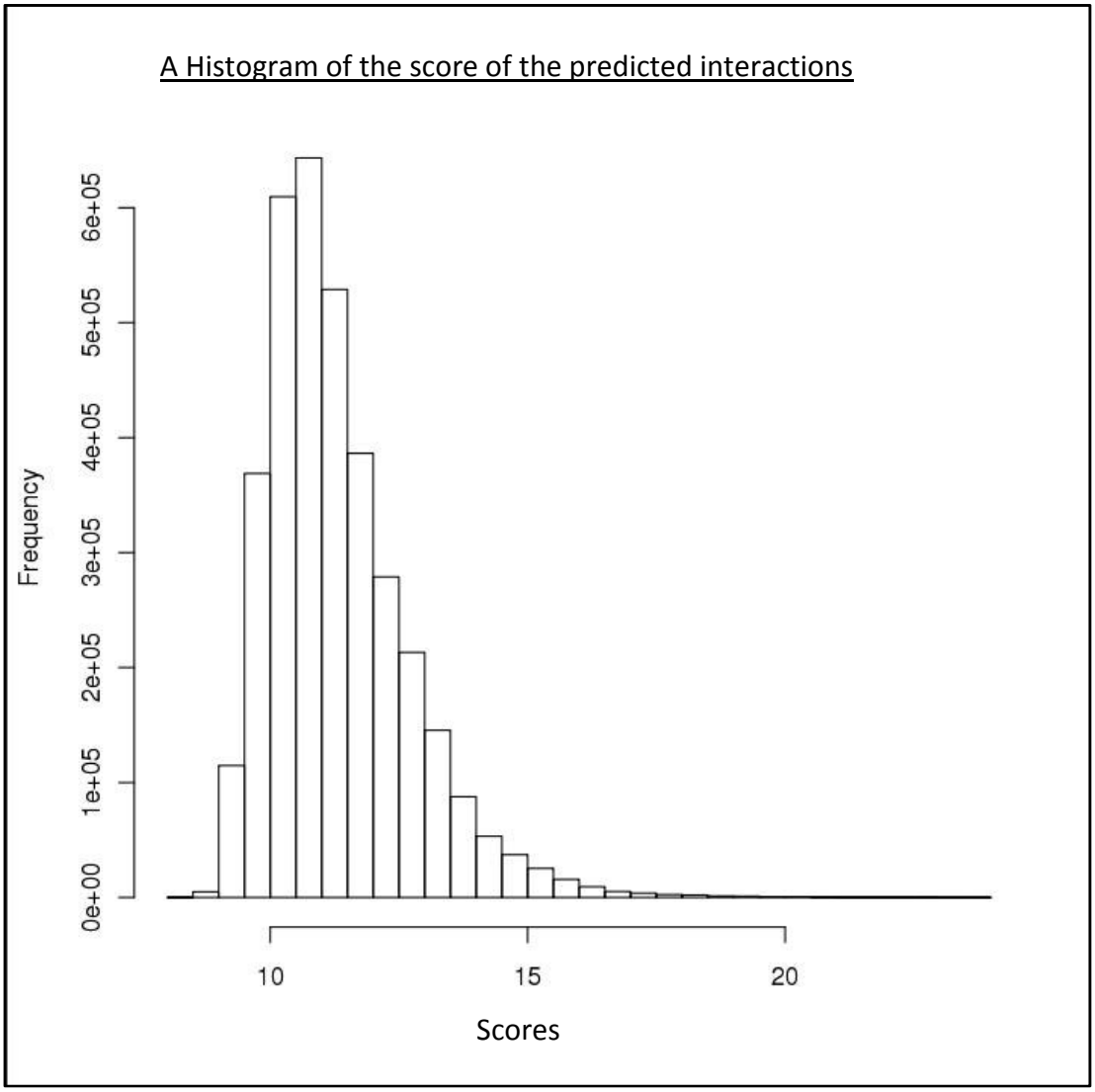

Figure 2.28 A histogram of the scores of all the predicted host-pathogen protein-protein interactions

2.5 Discussion

The goal of DISCOVERY is to integrate various relevant sources of annotation information about malaria. This should allow easier interpretation of malaria knowledge and may aid the discovery of novel drug targets. Host-pathogen interaction studies are a recent addition to drug discovery. Theoretically, accurate and reliable *in silico* host-pathogen protein-protein

interaction predictors should prove to be immensely important and could be applied as part of the framework of drug discovery studies.

This project, forming part of DISCOVERY, aims to predict accurate and reliable host-pathogen protein-protein interactions. Host-pathogen protein-protein interactions predictions were made using the protein data available in DISCOVERY and interaction data from DIP and MINT. Unique proteins were selected from the mentioned data and custom clusters were generated using ORTHOMCL. Integration of ORTHOMCL data, DIP and MINT data and data already available in DISCOVERY respectively, enabled *in vitro* host-pathogen protein-protein interactions predictions. DISCOVERY attempts to apply theoretical weights based on sequence similarity, PEXEL / VTS motif presence, metabolic map sharing, microarray expression fluctuation and sub-cellular location to identify possible *in vivo* host-pathogen protein-protein interactions.

One of the greatest challenges in developing host-pathogen protein-protein interaction predictions is the lack of proven experimental interaction data. This challenge made host-pathogen predictions based on orthologs almost inevitable. The use of orthologs to make predictions enabled expansion of unrelated interaction data toward more specific malaria related data. Consequently, proteins shared in an ortholog group could be substituted with each other, aiding the derivation of host-pathogen interactions from existing protein-protein interactions obtained from interactions databases.

This project differs from the work of Lee *et al*. in various ways although the basic idea is the same. DISCOVERY uses custom made ortholog clusters generated by ORTHOMCL where Lee *et al*. uses a matrix approach to obtain ortholog clusters. ORTHOMCL was designed to handle malaria data specifically and might be a better approach than best reciprocal hit BLAST. ORTHOMCL also identifies paralogs and orthologs; incorporating recent only paralogs into the ortholog groups.

In contrast best reciprocal hit BLAST does not distinguish between recent and evolutionary paralogs.

Furthermore, DISCOVERY uses more analytic sources to identify possible *in vivo* host-pathogen interactions than any of the previous methods (Dyer, *et al*., 2007; Lee, *et al*., 2008). DISCOVERY's scoring rather than filtering method also makes it more sensitive than known predictors as it does not exclude any predictions. Predictions are simply rated according to a score varying from approximately 5 to 20. The calculation and weight assignment for possible *in vivo* interactions however, is challenging as weights and scores are assigned theoretically and little evidence is available to support scoring. Never-the-less, scoring should indicate *in vivo* occurrences as various different sources were incorporated into scores.

In the next chapter we attempt to assess accuracy and reliability using comparative studies.

Chapter 3: Analysis and comparison of predicted host-pathogen protein-protein interactions

3.1 Introduction:

It is crucial to test DISCOVERY's host-pathogen interactions prediction power against relevant known data. Interaction reliability and accuracy is usually tested with comparison studies. Unfortunately the serious lack of experimental interaction data makes extensive comparative studies with malaria host-pathogen interactions difficult, none the less interaction comparisons remain a useful strategy.

3.1.1 Experimental interactions

Host-pathogen interaction studies revealing actual protein-protein interactions between species are rather limited. Lovegrove *et al.* (2006) claims that only a small amount of such studies exist. Most of these host-pathogen interaction studies are based on *in vitro* and *in vivo* expression patterns of either a host or the *Plasmodium* during infection; meaning that the focus still remains one-sided. Interestingly, malaria host-pathogen interaction studies seldom focus on the host-pathogen interactions between a human host and *Plasmodium* parasite. Literature searches exposed only one such method, that reveals actual protein-protein interactions between *H. sapiens* and *P. falciparum* (Vignali*, et al.*, 2008). Other malaria-based host-pathogen interaction studies focus on interactions between the *Plasmodium* parasite and *Anopheles,* the vector-host.

3.1.1.1 Specific interactions between *Plasmodium* and *Anopheles*

Despite a multitude of information available on the role of *Anopheles* during transmission of *Plasmodium,* little is known about the biological processes and interactions taking place

between these species during the developmental cycle of gametes to sporozoites. A handful of studies focus on this problem, a discussion of two of these studies follows.

Basseri *et al*. studied the effects of carbohydrates on the development of *Plasmodium* gametocytes to sporozoites in *Anopheles* mosquito. Their results recognized an inhibitory effect on sporozoite development with mosquitoes being fed a sugar supplemented, *Plasmodium* infected blood meal. Arabinose and fucose revealed an inhibitory interference on oocyst formation in the midgut of the *Anopheles*. Interestingly, no sporozoites were found in the salivary glands of *Anopheles* with mannose, GALNAc and lactose supplemented diets (Basseri*, et al.*, 2008). The last mentioned results revealed a promising inhibition effect as sporozoites weren't developed and therefore could not be transferred to a *H. sapien* host.

In a more direct approach Ghosh *et al.* utilized information from SM1, a twelve-amino-acid peptide that is known to bind to the *Anopheles* salivary gland and inhibit *Plasmodium* sporozoite invasion, to identify its target *Anopheles* protein. The detection of the *Anopheles* protein saglin as SM1's target protein was inferred using UV-crosslinking (Ghosh*, et al.*, 2009). Additionally an anti-SM1 antibody is used to identify a peptide mimotope TRAP from *Plasmodium*. Saglin was found to bind with TRAP. Further analysis on saglin illustrated that down regulation thereof leads to the inhibition of salivary gland invasion. This implies that the saglin / TRAP interaction plays a vital role during *Plasmodium* invasion. The invasion stage of *Plasmodium* species is crucial for parasite survival and infection. This makes saglin and TRAP key drug targets.

DISCOVERY does predict interactions between *A. gambiae* and *P. falciparum*, but this comparison study aims to only compare interactions between *H. sapiens* and *P. falciparum*. Therefore, comparisons between *A. gambiae* and *P. falciparum* will not be discussed.

3.1.1.2 Interactions between *H. sapiens* and *P. falciparum*

Vignali *et al*. made use of a modified yeast two-hybrid approach described previously (LaCount, *et al*., 2005). This process preferentially selects plasmids that contain fragments of proteins that can be expressed in yeast.

The DNA-binding domain (BD) library is constructed from *P. falciparum* proteins that are expressed during the IDC, whereas the activation domain (AD) is constructed to include mRNA from *H. sapiens'* liver and cerebellum (Vignali, *et al*., 2008). Like all yeast two-hybrid methods Vignali *et al.* identified various false positive interactions from their results. Over-represented putative interactions were filtered out. Annotated and predicted GO annotations were used to identify interactions that may not occur *in vivo*, these false positives were also filtered out. After filtering most of the false positives, Vignali *et al.* selected interactions that are likely to occur during pathogenesis (true positives) as possible host-pathogen interactions.

The initial set of interactions of 2200 was carefully curated to find 456 relatively reliable interactions. After inspection of these interactions Vignali *et al.* exposed a cluster where the *P. falciparum* protein PFE1590w/ETRAMP5 interacts with human apolipoproteins ApoA, ApoB and ApoE. ETRAMP5 is a parasitophorous vacuole membrane protein and alolipoproteins are known to transport dietary fats through the bloodstream. After further investigation Vignali *et al*. established that an ApoE genotype affects the risk for malaria infection. Other than ETRAMP5 other possible drug targets were also determined (Vignali, et al., 2008).

The above mentioned 456 interactions were used in comparisons with other known *in silico* predictors' results. A discussion on the comparison results will follow in Section 3.3.

3.1.2 *In silico* interactions

Literature searches revealed only 2 *in silico* methods that aim to predict host-pathogen protein-protein interactions between *P. falciparum* and *H. sapiens*. These methods will now be discussed.

3.1.2.1 Bayesian statistics as an approach for host-pathogen interactions by *Dyer et al*

Dyer *et al.* suggests the use of protein domains to determine protein interactions between different species (Dyer, *et al*., 2007). Dyer *et al*'s predictions are based on Sprinzak *et al.*'s sequence-signature algorithms, with the only difference being that *Dyer et al.* uses it to identify interspecies protein-protein interactions (Sprinzak and Margalit, 2001). The sequence-signature algorithm aims to recognize recurring sequence characteristics from experimental data. Identification of recurring sequence signatures between interacting proteins enables the structuring of a network of relationships. This network of different sequence signatures in turn could be used to examine and predict other possible protein-protein interactions. Dyer *et al*. makes use of UNIPROT (Apweiler, *et al*., 2004) data and domain data from INTERPRO-SCAN (Quevillon*, et al.*, 2005) to construct this network.

Bayesian statistics are applied on the domain pairs identified by the sequence signature algorithm. The probabilities determined *via* Bayesian statistics are used to examine the likelihood of interactions occurring between host and pathogen proteins. These statistics will now be discussed in more detail.

3.1.2.1.1 Statistics

Let D(g, d) be the event where protein g contains domain d. Furthermore let I(g, h) be the event where protein g interacts with protein h. Taking these events into account, Bayes rule

can be applied to determine the probability of protein g interacting with protein h given that protein g contains domain d and protein h contains domain e (Equation 3.1).

Equation 3.1 Bayes Rule

$$\Pr\{g,h \mid d,e\} = \frac{\Pr\{d,e \mid g,h\}\Pr\{I(g,h)\}}{\Pr\{D(g,d)\}\Pr\{D(h,e)\}}$$

Now, let P be the event where a protein contains at least one domain and takes part in a protein-protein interaction. From this, let S contain the set of interaction pairs from event P. Concurrently, S(d, e) will be the event where interaction pairs contain both domain d and domain e respectively. Using this information the probability that domain d interacts with domain e, given that domain d exists within protein g and domain e exists within protein h, as well as the probability that protein g interacts with protein h and the probability that protein g contains domain d and protein h contains domain e could be calculated according to Equation 3.2.

Equation 3.2 Bayes Rule factors

- $\Pr\{d,e \mid g,h\} = \frac{|S(d,e)|}{|S|}$
- $\Pr\{I(g,h)\} = \frac{|S|}{\binom{|P|}{2}}$
- $\Pr\{D(g,d)\}, \Pr\{D(h,e)\} = \frac{|P_d||P_e| - (P_d \cap P_e)}{\binom{|P|}{2}}$

Each of the factors from Equation 3.2 could then be substituted into Equation 3.1, to get Equation 3.3.

Equation 3.3 Substitution of Bayes factors into Bayes Rule

$$\Pr\{g,h \mid d,e\} = \frac{|S_{d,e}|}{|P_d \parallel P_e| - |P_d \cap P_e|}$$

In cases where multiple domain pairs predict the same protein pairs to interact, data are considered independent of each other. After each interaction's probability were calculated according to Equation 3.3, interaction pairs with a probability lower than 0.5 were discarded from further analysis. Interaction pairs with a probability higher than 0.5 were classified as possible host-pathogen protein-protein interactions.

3.1.2.1.2 Three tests to analyze the predicted host-pathogen interactions

After the host-pathogen protein-protein interactions were predicted, *Dyer et al.* constructed a few tests to analyze the interactions.

3.1.2.1.2.1 Proximity in interspecies

The first step of proximity testing is to identify and group the predicted interactions into triplets where pairs of host proteins interact with the same pathogen protein. A distance measure is then determined between the host pairs using existing protein-protein interaction networks. After the distances between all the H-H-P triplet interactions were determined, a distribution distance was calculated for each pair over the distances of all the interaction pairs. Similar to the host distribution distances, pathogen distribution distances were calculated for pairs of pathogen proteins that interacted with the same host protein.

The principle of proximity is that proteins that are in close proximity to each other are more likely to co-interact with the same target protein. Therefore, predicted interactions were

weighted according to their proximity between either the host protein pairs or the pathogen protein pairs.

3.1.2.1.2.2 Gene expression

According to various different articles (Grigoriev, 2001; Jansen*, et al.*, 2002), proteins that interact within the same organism share expression profiles. This principle originates into the second test formulated by *Dyer et al*.

Expression profiles from the various stages of the parasite lifecycle were used to enrich previously defined HHP (host-host-pathogen) and PPH (pathogen-pathogen-host) triplets. Each pair within the triplets is checked for correlation of expression profiles using Spearman's coefficient correlation. Afterwards, these correlations are plotted with similar distribution for all the triplets.

3.1.2.1.2.3 GO annotation

The host-pathogen protein-protein interactions were further enriched with functional GO categories where possible.

3.1.2.1.3 Filtering

After host-pathogen predictions were made certain filtering of data was necessary to retrieve only the protein data of proteins that are active during infection.

Firstly, *Plasmodium* proteins that were annotated with mitochondrion, nucleus, ribosome, cellular processes, helicase activity, complex activity, nuclease activity, nucleic binding, nucleotide binding were excluded from protein data. Furthermore all the *H. sapiens* proteins

annotated with ribosome, nucleic, nucleic binding, nucleoside binding or proteolysis activity were also excluded. After excluding all these proteins, *Plasmodium* proteins annotated with subtilisin activity, dense granule, hemoglobin metabolism, protein folding, polymerization, cell-cell communication or cell death as well as *H. sapiens* proteins and *Drosophila* proteins annotated with blood coagulation, cell-cell communication, protein folding, polymerization or cell death were again included into protein data even if it was previously excluded. Thirdly, all the proteins not taking part in any protein-protein interactions were excluded. Finally, all the proteins with little evidence of interacting domains were excluded from predictions.

3.1.2.1.4 Results

Dyer et al. predicts 516 protein-protein interactions between *H. sapiens* and *P. falciparum*. Their results showed that host protein pairs that interact with the same pathogen are distantly located to each other according to gene maps. Additionally, *Plasmodium* protein pairs that are predicted to interact with the same *H. sapiens* protein are co-expressed in microarray datasets.

The predictions from Dyer *et al*. (528 interactions) are compared with both DISCOVERY and Lee *et al.'*s results in section 3.3.

3.1.2.2 Ortholog-based protein-protein interactions prediction for interspecies by *Lee et al*

Similar to DISCOVERY, *Lee et al*. also use an ortholog-based approach to predict host-pathogen protein-protein interactions. An ortholog-based approach exploits previous knowledge about conserved interactions shared in different species to construct interaction networks for specific proteins. *Lee et al*. obtained this data from POINT (Huang*, et al.*, 2004), a functional database previously developed by their group. POINT combines information from various public interaction databases with the focus of extracting mouse, fruitfly, worm and yeast proteins and

converting them to a predicted human interactome. This conversion step is based on the idea of conserved orthologs. In addition to predictions, POINT includes expression data from microarrays and ontology data from GO to improve the spatial proximity of interactions.

Lee et al. takes the general usage of orthologs to predict protein-protein interactions networks in a single organism a step further by applying the same principal to predict protein-protein interactions between two species. Although this application seems simple it complicates the identification of *in vivo* interactions considerably, since characteristics for both species must now be taken into account. A single factor like spatiotemporal constraints already complicates matters.

In summary Lee *et al*. first isolated the proteins from 18 eukaryotic species and experimental proteins from HOMOLOGENE (Sayers, *et al*., 2009) and POINT (Huang, *et al*., 2004) respectively. Integration of this data was used to construct host-pathogen protein-protein interactions.

3.1.2.2.1 Construction of ortholog matrix

Lee *et al.* used the 18 eukaryotic species' proteins extracted from HOMOLOGENE (Sayers, *et al.*, 2009) to construct an ortholog matrix. The *H. sapiens* and *P. falciparum* species are included in these 18 eukaryotes. This matrix served as a deterministic tool; identifying orthologs of higher eukaryote organisms. Analysis of this ortholog matrix established that 81 of all the genes were conserved in all 18 species. Strikingly, 243 of the genes present in *P*. *falciparum* did not exist in any of the other 17 species.

3.1.2.2.2 Extracting protein-protein interactions from POINT

The protein-protein interaction data from POINT (Huang, *et al*., 2004) was used to identify interlogs in *H. sapiens* and *P. falciparum*. Most of the interactions in POINT were gathered from high-throughput techniques like yeast two-hybrid. High-throughput techniques are likely to produce lots of false positive interactions. Therefore, Lee *et al*. decided to use only the interactions confirmed by two or more different techniques and interactions that occurred in more than two articles for further predictions.

3.1.2.2.3 Inferring interlogs from the ortholog matrix

After selecting more reliable protein-protein interactions from POINT, filtering was continued. Using the ortholog matrix, interlogs could now be inferred. An interlog met Lee *et al.*'s basic requirements if it was supported by having at least two conserved orthologs in different species. Interlogs that did not meet these requirements were excluded from further analysis.

According to Lee *et al*'s analysis of the ortholog matrix, only 990 of *P. falciparum's* proteins shared orthologs in one of the 17 species. When taking this knowledge into account, interlog filtering seems too stringent. Only 990 out of the total number 5266 genes were orthologs conserved over different species. This meant that at most 20% of *P. falciparum*'s genes were included in Lee *et al.*'s current representation of interlogs. It could be argued that the 80% of genes that are excluded from further analyses might be the exact genes that pertain to the *P. falciparum* parasite's unique abilities to evade and interact with the human-host proteins.

3.1.2.2.4 Prediction of host-pathogen protein-protein interactions

Integration of the interlogs and filtered protein-protein interactions led to the prediction of 3090 *in vitro* interactions between *H. sapiens* and *P. falciparum*. Similar to DISCOVERY, these *in*

vitro interactions needed to undergo a filtering process before *in vivo* host-pathogen interactions could be retrieved.

This was accomplished using the Gene Ontologies (Go) of *H. sapiens* and *P. falciparum* proteins. The proteins taking part in host-pathogen protein-protein interactions were characterized according to Go biological processes (Figure 3.1).

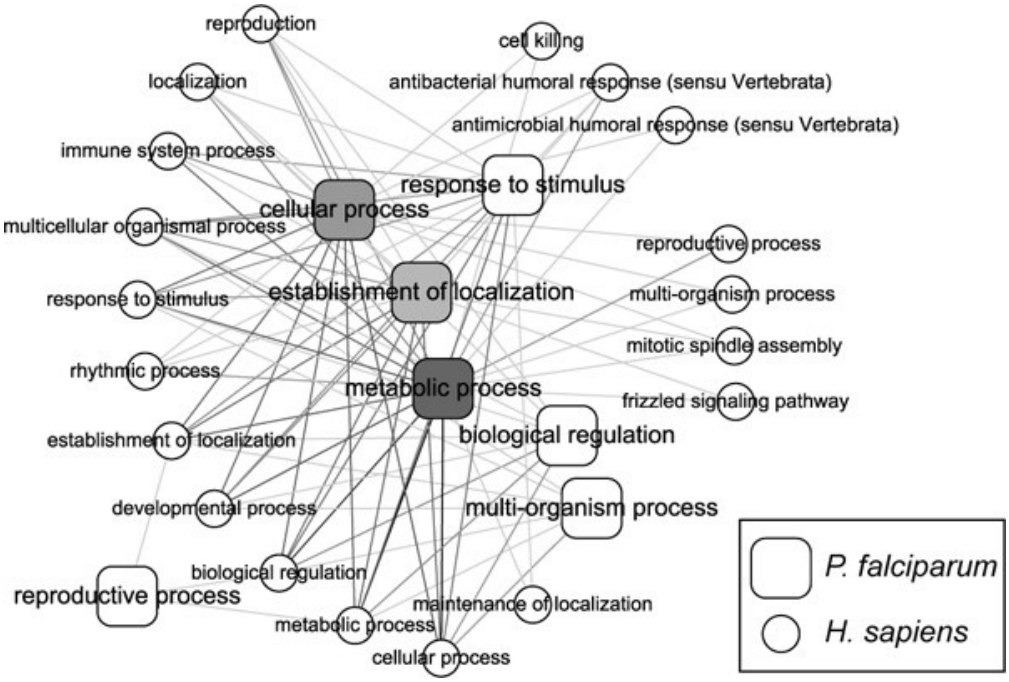

Figure 3.1 Host-pathogen protein-protein interactions grouped according to biological processes (Lee, *et al.*, 2008)

According to Figure 3.1 most host-pathogen interactions occur between metabolic and cellular processes. Although this phenomenon can be described as *P. falciparum*'s dependency on host cells' nutrients for survival (Date and Stoeckert, 2006), the question of the stringency on interlog retrieval again comes to mind.

Out of this 3000 ontology enriched host-pathogen protein-protein interactions only 918 interactions remained after taking into account spatiotemporal constraints. This was followed by filtering of host-pathogen protein-protein interactions using the presence/absence of translocation signals; only 95 interactions remained.

Within these 95 interactions, *Lee et al.* confirmed one *P. falciparum* protein which interacted with 50 *H. sapiens* proteins. These 50 *H. sapiens* proteins had calmodulin like functions. Although *Lee et al*. ended up with 95 host-pathogen protein-protein interactions, the interactions from supplementary file 4 (3090 interactions) was used for comparison studies between Lee *et al.,* Dyer *et al*. and DISCOVERY.

3.2 Methods

The first step towards comparisons between DISCOVERY's host-pathogen predictions and the results from Dyer *et al.* and Lee *et al.* was to retrieve all the unique interaction predictions between *H. sapiens* and *P. falciparum* from DISCOVERY.

As mentioned earlier DISCOVERY's predictions are based on proteins extracted from ENSEMBL and PLASMODB. Closer inspection revealed that Dyer *et al*'s results were based on gene names and UNIPROT protein ids and Lee *et al*'s results on gene names and ENSEMBL protein coding ids. To enable comparisons between these different studies, their ids were transformed to curated, non-redundant UNIPROT protein ids. During the transformation process some ids could not be linked to UNIPROT proteins, interactions containing these ids were discarded. Transformation of the different studies' ids could be divided into a few steps:

- DISCOVERY
    - An ENSEMBL id or PLASMODB id has existing links to a UNIPROT alias, if no alias existed the interaction containing the ids was discarded
- Dyer *et al*.
    - Find all the UNIPROT ids that co-occur in DISCOVERY's predictions
    - Find UNIPROT ids (www.uniprot.org) linked to the gene names, and then make sure the UNIPROT ids are the current ids in use
- Lee *et al.*
    - Directly link ENSEMBL ids to DISCOVERY and then find its UNIPROT alias
    - Convert gene names to UNIPROT ids with the same procedures followed with Dyer *et al,*'s results

The difference in publication dates between the studies suggested that some of the results may contain outdated UNIPROT protein ids, UNISAVE (refer to Section 2.2.2) was used to convert the older ids to recent ids.

The same process was repeated with experimental results. After having all the results in a comparable format, their interactions were checked against the NEGATOME database (refer to Section 2.3.6). No hits were found, therefore no further results were discarded and comparisons could be done.

3.3 Comparison results

Comparative studies are often used to measure a method's accuracy and confidence in results. Comparisons usually occur between known results and a method's results that need to be

evaluated. Unfortunately host-pathogen interaction studies in malaria lack the comprehensiveness of a source of proper accurate data.

The comparisons between the above mentioned *in silico* results with experimental results and comparisons between the results of the different *in silico* methods with each other follows.

3.3.1 Comparing experimental results with *in silico* results

Vignali *et al.*'s results were compared with interaction predictions from Lee *et al.,* Dyer *et al*. and DISCOVERY. This comparison revealed no similar interactions between any of the *in silico* predictors' results and Vignali *et al.*'s yeast two-hybrid results.

3.3.2 Comparisons between *in silico* methods

As illustrated in Vignali *et al.*'s results, host-pathogen interaction studies in malaria lack the comprehensiveness of a source of proper and accurate data. Although no comparisons could be made between *in silico* and empirical methods, it remains important to compare *in silico* methods with each other in an attempt to identify consensus-based interactions. *In silico* methods are limited, but additionally, comparing DISCOVERY with Lee *et al*. and Dyer *et al*. may give some measure of accuracy and confidence in DISCOVERY's predictions.

A crucial step in comparing *in silico* method results is to establish a common protein id source. It was decided to translate the protein ids of the *in silico* results to UNIPROT format. Since UNIPROT is a non-redundant database, unnecessary duplicate predictions would be excluded from the comparisons. Over time it often happens that UNIPROT ids are updated or replaced by new ids, UNIPROT tracks these changes with a database called UNISAVE (Leinonen, *et al.*, 2006). As far as possible, all predicted interaction's proteins were translated to UNIPROT format; if only

one of the proteins in an interaction could be translated the interaction was discarded from further comparison purposes. The format of the proteins in the *in silico* methods vary from UNIPROT gene names, ENSEMBL ids and UNIPROT ids. Gene names are searched for against the SWISSPROT database and then the SWISSPROT ids are returned. If the ids could not be found in the local SWISSPROT databases included in DISCOVERY, they were searched online via the UNISAVE webpage. ENSEMBL ids are translated using the alias columns in the *protein* table of DISCOVERY. After translation, the data of each method was stored into the database. Thus, comparison results were made readily available with a simple SQL query.

DISCOVERY results yielded a total number of 47,604 possible interactions between *P. falciparum* and *H. sapiens* (scores between approx. 5 and 20). After translating *Lee et al.* and *Dyer et al.* to UNIPROT format they respectively yielded *2*943 and 339 interactions. This meant a loss of 148 and 177 interactions respectively. This loss of interactions can be explained by investigating the format of the original data; ENSEMBL contains redundant entries and in some cases no UNIPROT ids could be linked to gene name ids.

Figure 3.2 show the total number of comparisons present between the different *in silico* prediction methods.

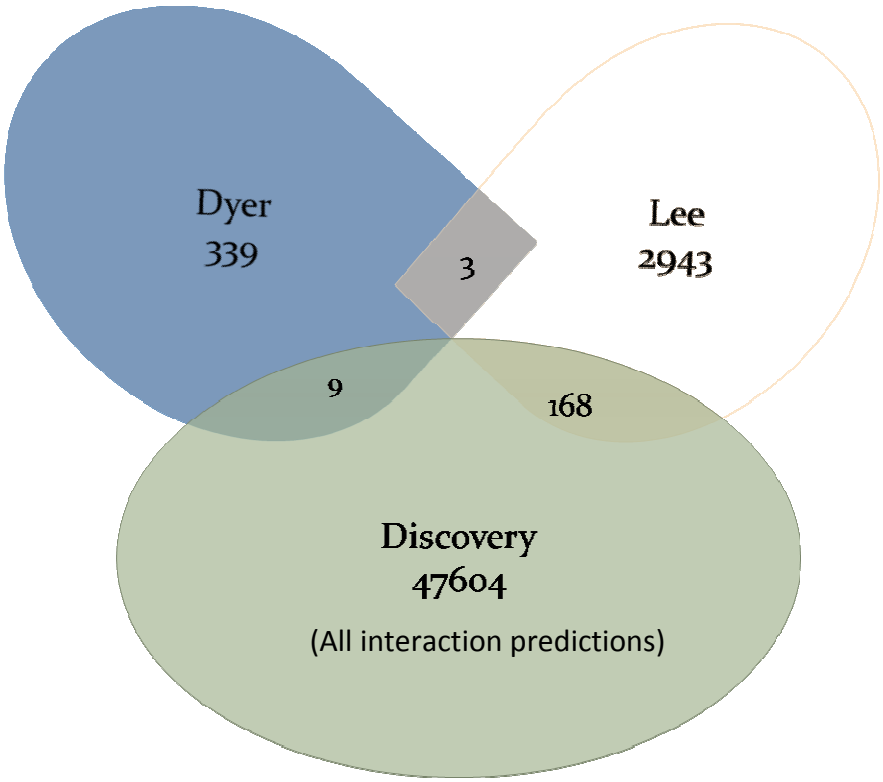

Figure 3.2 Comparisons between *in silico* methods, all of DISCOVERY'S interactions were compared with Dyer *et al* and Lee *et al's* interactions

The comparative results were analyzed and evaluated.

3.3.2.1 *Dyer et al.* vs DISCOVERY

Consequently, all 9 interaction similarities between DISCOVERY and *Dyer et al.* revealed participating pathogen proteins involved in the pyruvate dehydrogenase complex (PDC). Figure 3.3 illustrates the PDC mechanism. The PDC mechanism exists of three subunits, pyruvate dehydrogenase (E1), dihydrolipyol transacetylase (E2) and dihydrolipyol dehydrogenase (E3). Five out of the nine interactions contained the dihydrolipoamide acyltransferase E2 as pathogen protein; the remaining interactions contained the pyruvate dehydrogenase E1 beta subunit as pathogen protein.

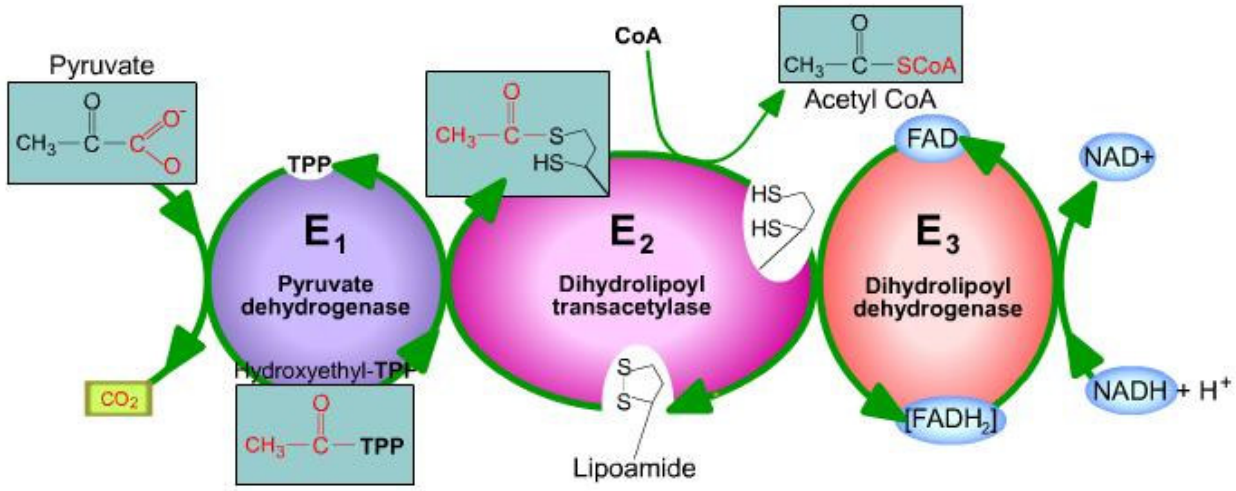

Figure 3.3 The mechanism of the pyruvate dehydrogenase complex (PDC), the PDC acts as a mediator between glycolysis and the citrate acid cycle (http://www.wiley.com/college/boyer/0470003790/animations/pdc/pdc.htm)

The PDC is a large mutli-enzyme complex located within the mitochondria. This complex is responsible for enhancing oxidative decarboxylation, acetyl unit transfer to coenzyme A (CoA) and eventually the reduction of NAD+ to NADH by means of electron uptake (Guan*, et al.*, 1995). The three enzymes pyruvate dehydrogenase (E1), dihydrolipoamide dehydrogenase (E2) and dihydrolipoamide dehydrogenase (E3) form the PDC structure. These enzymes consecutively collaborates, **E1** is responsible for decarboxylation of pyruvate and reductive acylation of lipoyl moiety which in turn covalently bonds with **E2**. **E2** transfers acyl to CoA and **E3** reoxidize the oxidized lipoamides (Guan, *et al*., 1995).

According to the predictions, proteins interacting with E1 vary between pyruvate dehydrogenase protein X, E2 chains and dihydrolipoyllysine residue components of pyrovate dehydrogenase E1. Furthermore, proteins are predicted to interact with E2 are E1 and 2-oxoisovalerate dehydrogenase, a branched-chain of the alpha-keto dehydrogenase.

In most organisms these interactions occur independently and the proteins necessary for these processes are present. It is therefore very unlikely that these kinds of interactions will take place between different species. DISCOVERY interaction scores for these interactions varies from 8-14, when referring to Figure 2.28, the score of 14 seems high. It might therefore be considered to adjust the weights to improve DISCOVERY's discriminating power.

Comparing Dyer *et al.'s* results with DISCOVERY'S results allow little room for speculation on a threshold score that can discriminate between *in vitro* and *in vivo* interactions. Investigation of the interactions that were present in both Dyer *et al.* and DISCOVERY revealed interactions that are able to occur without interference from a host or pathogen, since both of the interacting proteins are present in each individual species.

3.3.2.2 *Lee et al.* vs DISCOVERY

According to Figure 3.2, Lee *et al*. and DISCOVERY shares 168 comparisons. It is imperative to identify only *in vivo* interactions before analyzing these results. After investigating the results of Dyer *et al*. a threshold of 15 was used to analyze results between Lee *et al*. and DISCOVERY.

This analysis' results are shown in Figure 3.4.

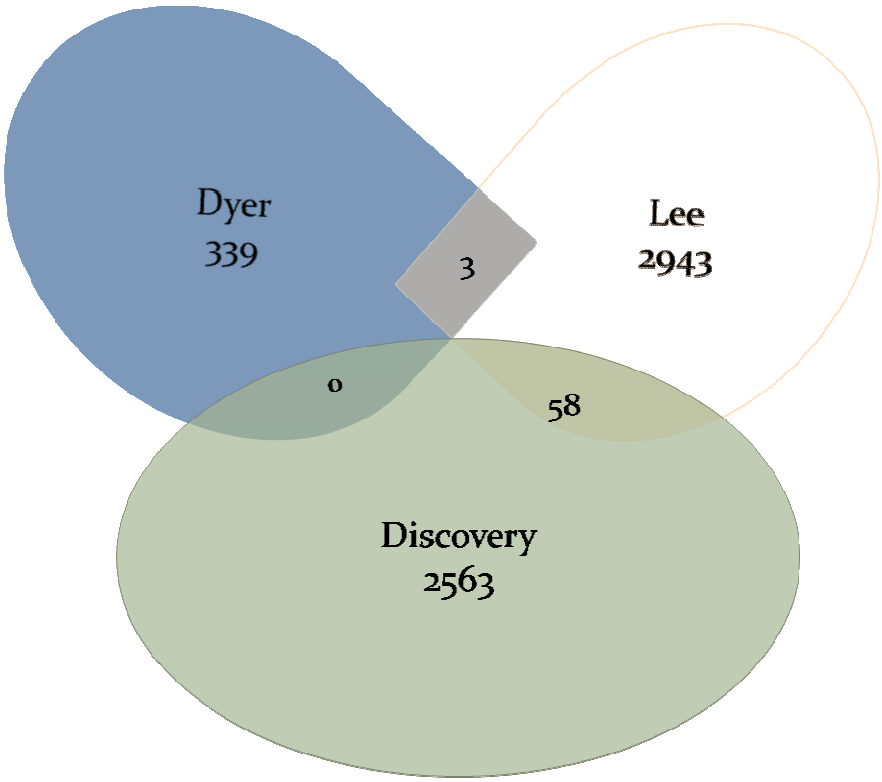

Figure 3.4 Comparison results after applying a threshold score of 15

For illustrative purposes, a few of these interactions were selected and analyzed. Table 3.1 shows the results; an e-value threshold of 1×10^{-50} was used to ensure high specificity.

Table 3.1 Blast results with a threshold of 1 x 10^{-50}

With a threshold of $1e^{-50}$					
Pathogen			**Host**		
Uniprot id	Gene name	Species	Uniprot id	Gene name	Species
Q8IKV9	PF14_0492	*P. falciparum*	Q08209	PPP3CA	*H. sapiens, P. falciparum* ~ serine /threonine protein phosphatase
Q8IBA0	PfRACK	*P. falciparum, H.sapiens* ~ Lung cancer oncogene	P12931	SRC	*H. sapiens*
Q7KQK2	pUB	*P. falciparum, H.sapiens* ~ Ubiquitin C	P04637	TP53	*H. sapiens*
Q8ILW9	PF14_0124	*P. falciparum, H.sapiens* ~ Actin	Q9UQ03	CORO2B	*H. sapiens*
Q8IIR9	CK2alpha	*P. falciparum, H.sapiens* ~ Casein kinase II	P67870	CSNK2B	*H. sapiens*

The confidence scores of the interactions in Table 3.1 varied between 16.5 and 20. DISCOVERY'S predictions successfully identify *in vivo* interactions. One question remains unanswered; are these *in vivo* interactions dependent on both host and pathogen species to occur, or do these interactions occur independently in individual species only? This question is not answered in the current project, but it is an eminent barrier that needs attention.

3.4 Interactions predicted by DISCOVERY

Further analyses of interactions lead to the question of evidence of *in vivo* interactions that were identified to be dependent on two different species. A quick search revealed very little such predictions, but a previously identified interaction between C6KT34 and Q9UNZ2 is one such example.

C6KT34 is annotated as a cell division cycle protein from *Plasmodium* that belongs to the AAA ATPase family, whereas Q9UNZ2 is annotated as a NSFL1 cofactor p47 (Uniprot). NFSL1 may reduce the ATPase activity of the valosin-containing protein (VCP). It is essential for the fragmentation of golgi stacks during mitosis and for VCP-mediated reassembly of golgi stacks after mitosis. Furthermore Q9UNZ2 might play a role in VCP-mediated formation of transitional endoplasmic reticulum (tER) according to similarity and is also known to inhibit cathepsin (CTSL) activity *in vitro*. CTSL is responsible for the overall degradation of proteins in lysosomes. Thus, if this interaction does take place *in vivo,* golgi fragmentation and protein degradation may be influenced. These kinds of interactions may be involved in parasite evasion strategies.

Unfortunately no evidence of this interaction has been found in literature studies, which again highlights the lack of available host-pathogen interaction data.

3.5 Discussion

The lack of knowledge about protein-protein interactions between different species makes host-pathogen interaction analyses a difficult task. Reliable experimental host-pathogen interaction results are crucial for comparison studies and the determination of the quality of *in silico* prediction methods.

According to our analysis *in silico* prediction methods compare poorly with experimental methods like yeast 2-hybrid screening. Bad comparability between experimental methods and *in silico* methods is a great hurdle to overcome. Although DISCOVERY's comparisons are briefly discussed, it is important to note that none of the *in silico* methods had any comparability to Vignali *et al*'s experimental results.

Various explanations can be given for the unanticipated poor comparability between DISCOVERY and Vignali *et al*'s results. Firstly, DISCOVERY's host-pathogen predictions are ortholog-based and therefore predictions are information-dependent. The main reason for the use of ortholog-based approaches is to overcome the challenge of limited gene/protein information. One difficulty might be that the data used to fill these gaps is poorly compatible with the species of interest or with the interactions between specific species. Secondly, DISCOVERY applies a filter over both the DIP and MINT interaction datasets to maintain the quality of the data that is used to make predictions. The quality filter applied by DISCOVERY is quite stringent and information about interactions that actually occur during malaria infections might be lost. Thirdly, continuously changing protein ids and proteins from different origins makes interaction comparisons between different studies a difficult and tedious task. Interaction information is often lost during the comparison process.

On the other hand, if the focus is shifted to Vignali *et al*. the first discrepancy would be the data used for the analyses. Vignali *et al*. use data that are expressed during the IDC stages of infection, where ortholog prediction tools would cover the whole process of infection. This high stringency on starting data of experimental studies is mainly due to cost effectiveness and time constraints of a study. After retrieving the interactions, these interactions are also filtered according to over represented interactions and annotated GO ontologies. This attempt to focus on *in vivo* interactions makes Vignali *et al's* techniques even less sensitive, and true positives might already be excluded after this process. After filtering out possible false positives *Vignali et al.* take it one step further, trying to select only the interactions that seems more likely to occur (true positives). In doing so even more possible interactions are excluded from their results. All that filtering makes Vignali *et al*'s approach very stringent, but it does not explain why none of the *in silico* methods share any interaction. It could be concluded that the initial data used for each of the host-pathogen interaction prediction methods are to divergent from each other and therefore their results differ so considerably.

In this Chapter, experimental results and two different *in silico* methods' results were compared with DISCOVERY's predictions. Similar to other *in silico* methods, DISCOVERY did not compare well with the experimental data (Doolittle and Gomez, 2010). Comparisons between the *in silico* methods first of all, revealed little interactions overlapping between Dyer *et al.* and Lee *et al*. Comparing DISCOVERY with Dyer *et al.* also had limited success, but Lee *et al.* and DISCOVERY had better success. The reason for bad comparability with Dyer *et al.* might be their use of different starting data and Dyer *et al.*'s unique approach of identifying interactions. Lee *et al.* and DISCOVERY's techniques are quite similar; both methods are based on predictions via orthologs. DISCOVERY differs from Lee *et al.* in that it uses ORTHOMCL to cluster orthologs and that it uses a broader spectrum of information to score possible *in vivo* host-pathogen interactions. The scoring of interactions, rather than filtering out interactions allows a user to specify their own threshold score when viewing interactions. Bridging the gap of experimental data could make *in silico* host-pathogen interaction predictions a key factor during drug discovery studies.

Chapter 4: Concluding discussion

Recent revisions on host-microbe interaction studies led to the conclusion that the current framework in use was ineffective. Definitions of the terms used to describe pathogenicity did not hold true, as definitions were based on an under-representative framework. Rather than accusing just microbes for pathogenicity, a new framework suggests that both microbes and hosts play an important role during infection / disease.

Previous studies have proven that protein functions could be correctly inferred from homology (Emes, 2008). Two kinds of homologs exists namely orthologs and paralogs, the assumption behind homology is that two proteins that have a common ancestor and have evolved quite recently, should have a similar function. If proteins share the same function their structural properties should be the same (Shin*, et al.*, 2007), meaning that these proteins might interact with similar partners. According to the above mentioned assumptions *in vitro* host-pathogen interactions (HPIs) can be predicted using a simple species filter.

Malaria causes over a million worldwide deaths per year. A proportion of 90% of these deaths occur in Sub-Saharan Africa. Currently no efficient vaccine for malaria exists, since *Plasmodium* continuously supersedes present drugs. The first step in drug discovery is to identify possible drug targets; the different *Plasmodium* species offer many promising drug targets. Therefore, implementation of computational methods to mine these drug targets is needed. Bioinformatics plays an important role in drug target discovery and drug design. Computer-based methods have the great advantage of easily integrating and evaluating of huge amounts of data, aiding scientists to study protein characteristics on genomic level. *In silico* HPI predictions only deliver *in vitro* interactions, consequently additional analyses are done to verify if interactions may occur *in vivo*.

According to literature studies various factors can be used to verify the likelihood of host-pathogen interactions occurring *in vivo*.

According to the assumption of homology of proteins, the closer the similarity between sequences the higher would the likelihood be of shared function. Predicted interactions' similarity to proven experimental interactions might give some power to the likelihood of the predicted interactions actually occurring either *in vivo* or *in vitro*.

The use of sub-cellular locations to define interactions as occurring *in vivo* is one of the more popular methods, but most of these applications occurred during determination within a single species. In such cases easy filtering on sub-cellular location sharing is possible, but not very sensitive. The determination of *in vivo* interactions between two different species becomes exponentially more difficult, never the less sub-cellular locations gives a good indication of *in vivo* interaction occurrence. Assignment of sub-cellular location scores according to inter-species interaction theory may still proof vital. Proteins that typically needs to pass more membranes or barriers are scored to be less likely to occur *in vivo,* than proteins adjacent to membranes or within cytoplasm.

Plasmodium relies on a parasitophorous vacuole (PV) to evade host erythorocytes. The evasion of host defense within erythrocytes is crucial, since expression profiles of *Plasmodium* proteins during this period suggest high metabolism and cell division activities (Bozdech, *et al*., 2003). This reproduction process occurring within erythrocyte is of utmost importance for pathogen survival. In order for metabolism processes to occur, the *Plasmodium* interacts with the host modeling some form of transport allowing resources to pass through the PV membrane (PVM). According to Horrocks *et al*. *Plasmodium* proteins contains a certain motif, called a PEXEL / VTS motif (Horrocks and Muhia, 2005) which allows protein to pass through the PVM. Proteins that

contain this motif are therefore more likely to interact with host proteins and interactions containing these pathogen proteins are therefore more likely to occur *in vivo*.

Microarray expression is used to verify possible protein-protein interactions (Hegde, *et al*., 2000). Although it remains unethical to study malaria in living humans, microarrays can still be used to identify the expression of *Plasmodium* proteins within cultured erythrocytes (Bozdech, *et al*., 2003). Using these results the expression of proteins could be measured by calculating the graph surface of a protein's expression during the intra-erythrocytic development cycle (IDC). A greater activity of *Plasmodium* proteins taking part in a predicted interaction makes the interaction more likely to occur *in vivo*.

Metabolic pathways are processes in which chemical compounds are modified to useable metabolites *via* a collection of enzymes. Therefore, if two proteins belong to the same collection of chemicals transforming a certain metabolite, these proteins are more likely to interact than random protein pairs. An analysis to determine if predicted host-pathogen protein-protein interaction proteins share a metabolic pathway should support the possibility of the proteins possibly interacting *in vivo*.

DISCOVERY is an integrative system; this integration of information from various sources increases the quality and power of DISCOVERY. DISCOVERY contain a collection of various data about malaria, the integration of these vastly differing sources gives DISCOVERY its power. The latest addition to this wealthy source of information is the prediction of possible *in vivo* host-pathogen protein-protein interactions. The analysis methods mentioned above were used to score interactions according to their likelihood of occurring *in vivo*. This addition in combination with the other information available in DISCOVERY will hopefully aid scientist in identifying possible drug targets for malaria.

DISCOVERY's host-pathogen interaction prediction tool has an advantage over known methods because it uses an ortholog clustering tool specifically created for malaria. Uniquely this tool identifies both orthologs and recent paralogs as orthlogs, since they are clustered together into orthologs groups. Another advantage is the use of more information to analyze possible host-pathogen interactions, which gives DISCOVERY a greater prediction power. According to DISCOVERY's results there also seems to be a down side to using orthologs to predict interactions between different species, as most of the predicted interactions seem to be able to occur independently within the individual species already.

One of the greatest challenges of predicting *in vivo* host-pathogen interactions is the **scoring or filtering step**, where *in vitro* interactions are identified as possible *in vivo* interactions. Scoring according to sub-cellular location between two different species proved to be quite a challenge as sub-cellular location sharing alone could not be used as a measure of *in vivo* occurrence of interactions. For instance in the case where two proteins that are predicted to interact are found in the nucleus, this finding actually makes an interaction very unlikely to occur *in vivo*. Another factor playing a role during sub-cellular location scoring is that any given protein's sub-cellular location was based on a prediction as well. Although the scoring characteristics seemed logical, the assignment of weights to integrate each characteristic into a final score has proven to be difficult as well. Ultimately, the lack of training data made it impossible to analyze DISCOVERY's scoring accuracy.

Similar to other host-pathogen predictions studies (Doolittle and Gomez, 2010), DISCOVERY did not compare well to yeast two-hybrid results. The same comparison between other known *in silico* predictors and yeast two-hybrid results, did not compare well either. This raises an important question on the verification of any *in silico* host-pathogen predictor. Overcoming this challenge of sufficient reliable and accurate experimental information will enable the

assessment of existing *in silico* host-pathogen interaction predictors. Surely, then the true power of host-pathogen interaction predictions will be realized.

The host-pathogen interaction leg of DISCOVERY shows a lot of promise, but is a project ahead of its time. Only when the lack of accurate experimental data is saturated the true potential of ortholog-based host-pathogen predictions will be discovered.

References

Accolla, R.S. (2006) Host defense mechanisms against pathogens, *Surg Infect (Larchmt)*, 7 Suppl 2, S5-7.

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

Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N. and Yeh, L.S. (2004) UniProt: the Universal Protein knowledgebase, *Nucleic Acids Res*, 32, D115-119.

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, 25-29.

Aurrecoechea, C., Brestelli, J., Brunk, B.P., Dommer, J., Fischer, S., Gajria, B., Gao, X., Gingle, A., Grant, G., Harb, O.S., Heiges, M., Innamorato, F., Iodice, J., Kissinger, J.C., Kraemer, E., Li, W., Miller, J.A., Nayak, V., Pennington, C., Pinney, D.F., Roos, D.S., Ross, C., Stoeckert, C.J., Jr., Treatman, C. and Wang, H. (2009) PlasmoDB: a functional genomic database for malaria parasites, *Nucleic Acids Res*, 37, D539-543.

Basseri, H.R., Doosti, S., Akbarzadeh, K., Nateghpour, M., Whitten, M.M. and Ladoni, H. (2008) Competency of Anopheles stephensi mysorensis strain for Plasmodium vivax and the role of inhibitory carbohydrates to block its sporogonic cycle, *Malar J*, 7, 131.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Sayers, E.W. (2009) GenBank, *Nucleic Acids Res*, 38, D46-51.

Bhardwaj, N. and Lu, H. (2005) Correlation between gene expression profiles and protein-protein interactions within and across genomes, *Bioinformatics*, 21, 2730-2738.

Blaschke, C., Hoffmann, R., Oliveros, J.C. and Valencia, A. (2001) Extracting information automatically from biological literature, *Comp Funct Genomics*, 2, 310-313.

Boddey, J.A., Moritz, R.L., Simpson, R.J. and Cowman, A.F. (2009) Role of the Plasmodium export element in trafficking parasite proteins to the infected erythrocyte, *Traffic*, 10, 285-299.

Bozdech, Z., Llinas, M., Pulliam, B.L., Wong, E.D., Zhu, J. and DeRisi, J.L. (2003) The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum, *PLoS Biol*, 1, E5.

Brown, K.R. and Jurisica, I. (2005) Online predicted human interaction database, *Bioinformatics*, 21, 2076-2082.

Cannon, W.B. (2009) *Encyclopaedia Britannica from Encyclopaedia Britannica 2007 Deluxe Edition*.

Caraco, T. and Wang, I.N. (2008) Free-living pathogens: life-history constraints and strain competition, *J Theor Biol*, 250, 569-579.

Casadevall, A. and Pirofski, L. (2001) Host-pathogen interactions: the attributes of virulence, *J Infect Dis*, 184, 337-344.

Casadevall, A. and Pirofski, L.A. (1999) Host-pathogen interactions: redefining the basic concepts of virulence and pathogenicity, *Infect Immun*, 67, 3703-3713.

Casadevall, A. and Pirofski, L.A. (2000) Host-pathogen interactions: basic concepts of microbial commensalism, colonization, infection, and disease, *Infect Immun*, 68, 6511-6518.

Chatr-aryamontri, A., Ceol, A., Palazzi, L.M., Nardelli, G., Schneider, M.V., Castagnoli, L. and Cesareni, G. (2007) MINT: the Molecular INTeraction database, *Nucleic Acids Res*, 35, D572-574.

Chen, F., Mackey, A.J., Vermunt, J.K. and Roos, D.S. (2007) Assessing performance of orthology detection strategies applied to eukaryotic genomes, *PLoS One*, 2, e383.

Clegg, R.M. (1995) Fluorescence resonance energy transfer, *Curr Opin Biotechnol*, 6, 103-110.

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

Datson, N.A., van der Perk-de Jong, J., van den Berg, M.P., de Kloet, E.R. and Vreugdenhil, E. (1999) MicroSAGE: a modified procedure for serial analysis of gene expression in limited amounts of tissue, *Nucleic Acids Res*, 27, 1300-1307.

Doolittle, J.M. and Gomez, S.M. (2010) Structural similarity-based predictions of protein interactions between HIV-1 and Homo sapiens, *Virol J*, 7, 82.

Drawid, A. and Gerstein, M. (2000) A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome, *J Mol Biol*, 301, 1059-1075.

Dyer, M.D., Murali, T.M. and Sobral, B.W. (2007) Computational prediction of host-pathogen protein-protein interactions, *Bioinformatics*, 23, i159-166.

Eisenhaber, F. and Bork, P. (1998) Wanted: subcellular localization of proteins based on sequence, *Trends Cell Biol*, 8, 169-170.

Emanuelsson, O. (2002) Predicting protein subcellular localisation from amino acid sequence information, *Brief Bioinform*, 3, 361-376.

Emes, R.D. (2008) Inferring function from homology, *Methods Mol Biol*, 453, 149-168.

EnzymeNomenclature Enzyme Nomenclature.

Fitch, W.M. (1970) Distinguishing homologous from analogous proteins, *Syst Zool*, 19, 99-113.

Förster, T. (1965) *Delocalized excitation and excitation transfer*. Academic Press Inc, New York.

Foster, S. and Phillips, M. (1998) Economics and its contribution to the fight against malaria, *Ann Trop Med Parasitol*, 92, 391-398.

Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.A., Copley, R.R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G. and Superti-Furga, G. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes, *Nature*, 415, 141-147.

Ghosh, A.K., Devenport, M., Jethwaney, D., Kalume, D.E., Pandey, A., Anderson, V.E., Sultan, A.A., Kumar, N. and Jacobs-Lorena, M. (2009) Malaria parasite invasion of the mosquito salivary gland requires interaction between the Plasmodium TRAP and the Anopheles saglin proteins, *PLoS Pathog*, 5, e1000265.

Gietz, R.D., Triggs-Raine, B., Robbins, A., Graham, K.C. and Woods, R.A. (1997) Identification of proteins that interact with a protein of interest: applications of the yeast two-hybrid system, *Mol Cell Biochem*, 172, 67-79.

Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., Vitols, E., Vijayadamodar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, A., Minto, M., Burgess, S., McDaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., Ioime, N., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carrolla, S., Bickelhaupt, E., Lazovatsky, Y., DaSilva, A., Zhong, J., Stanyon, C.A., Finley, R.L., Jr., White, K.P., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shimkets, R.A.,

McKenna, M.P., Chant, J. and Rothberg, J.M. (2003) A protein interaction map of Drosophila melanogaster, *Science*, 302, 1727-1736.

Gowda, M., Jantasuriyarat, C., Dean, R.A. and Wang, G.L. (2004) Robust-LongSAGE (RL-SAGE): a substantially improved LongSAGE method for gene discovery and transcriptome analysis, *Plant Physiol*, 134, 890-897.

Grech, K., Watt, K. and Read, A.F. (2006) Host-parasite interactions for virulence and resistance in a malaria model system, *J Evol Biol*, 19, 1620-1630.

Grigoriev, A. (2001) A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae, *Nucleic Acids Res*, 29, 3513-3519.

Guan, Y., Rawsthorne, S., Scofield, G., Shaw, P. and Doonan, J. (1995) Cloning and characterization of a dihydrolipoamide acetyltransferase (E2) subunit of the pyruvate dehydrogenase complex from Arabidopsis thaliana, *J Biol Chem*, 270, 5412-5417.

He, M., Wang, Y. and Li, W. (2009) PPI finder: a mining tool for human protein-protein interactions, *PLoS One*, 4, e4554.

Hegde, P., Qi, R., Abernathy, K., Gay, C., Dharap, S., Gaspard, R., Hughes, J.E., Snesrud, E., Lee, N. and Quackenbush, J. (2000) A concise guide to cDNA microarray analysis, *Biotechniques*, 29, 548-550, 552-544, 556 passim.

Henrici, A.T. (1934) Infection. In, *The biology of bacteria*. D.C. Heath and Co., Boston, Mass., 230-241.

Herman, B. (1999) Fluorescence Microscopy and Fluorescent Probes, Vol. 2, Edited by Jan Slavik 1998. Plenum Press, New York and London. 292 pages. (hardback, $95.00), *Microsc Microanal*, 5, 147.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A.R., Sassi, H., Nielsen, P.A., Rasmussen, K.J., Andersen, J.R., Johansen, L.E., Hansen, L.H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B.D., Matthiesen, J., Hendrickson, R.C., Gleeson, F., Pawson, T., Moran, M.F., Durocher, D., Mann, M., Hogue, C.W., Figeys, D. and Tyers, M. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry, *Nature*, 415, 180-183.

Horrocks, P. and Muhia, D. (2005) Pexel/VTS: a protein-export motif in erythrocytes infected with malaria parasites, *Trends Parasitol*, 21, 396-399.

http://expasy.org/enzyme/ Enzyme Nomenclature.

http://www.dpd.cdc.gov.dpdx Lifecycle of Malaria.

http://www.mysql.com MySQL.

http://www.sqlalchemy.org/ SQLAlchemy.

Huang, T.W., Tien, A.C., Huang, W.S., Lee, Y.C., Peng, C.L., Tseng, H.H., Kao, C.Y. and Huang, C.Y. (2004) POINT: a database for the prediction of protein-protein interactions based on the orthologous interactome, *Bioinformatics*, 20, 3273-3276.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin, E., Pettett, R., Pocock, M., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I. and Clamp, M. (2002) The Ensembl genome database project, *Nucleic Acids Res*, 30, 38-41.

Hulsen, T., Huynen, M.A., de Vlieg, J. and Groenen, P.M. (2006) Benchmarking ortholog identification methods using functional genomics data, *Genome Biol*, 7, R31.

Hurt, J.A., Thibodeau, S.A., Hirsh, A.S., Pabo, C.O. and Joung, J.K. (2003) Highly specific zinc finger proteins obtained by directed domain shuffling and cell-based selection, *Proc Natl Acad Sci U S A*, 100, 12271-12276.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome, *Proc Natl Acad Sci U S A*, 98, 4569-4574.

Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S. and Sakaki, Y. (2000) Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins, *Proc Natl Acad Sci U S A*, 97, 1143-1147.

Jackson, J.A., Pleass, R.J., Cable, J., Bradley, J.E. and Tinsley, R.C. (2006) Heterogeneous interspecific interactions in a host-parasite system, *Int J Parasitol*, 36, 1341-1349.

Jansen, R., Greenbaum, D. and Gerstein, M. (2002) Relating whole-genome expression data with protein-protein interactions, *Genome Res*, 12, 37-46.

Joubert, F., Harrison, C.M., Koegelenberg, R.J., Odendaal, C.J. and de Beer, T.A. (2009) Discovery: an interactive resource for the rational selection and comparison of putative drug target proteins in malaria, *Malar J*, 8, 178.

Joung, J.K. (2001) Identifying and modifying protein-DNA and protein-protein interactions using a bacterial two-hybrid selection system, *J Cell Biochem Suppl*, Suppl 37, 53-57.

Joung, J.K., Ramm, E.I. and Pabo, C.O. (2000) A bacterial two-hybrid selection system for studying protein-DNA and protein-protein interactions, *Proc Natl Acad Sci U S A*, 97, 7382-7387.

Kalyanaraman, C. and Jacobson, M.P. (2010) Studying enzyme-substrate specificity in silico: a case study of the Escherichia coli glycolysis pathway, *Biochemistry*, 49, 4003-4005.

Kim, W.K., Park, J. and Suh, J.K. (2002) Large scale statistical prediction of protein-protein interaction by potentially interacting domain (PID) pair, *Genome Inform*, 13, 42-50.

Kooij, T.W., Janse, C.J. and Waters, A.P. (2006) Plasmodium post-genomics: better the bug you know?, *Nat Rev Microbiol*, 4, 344-357.

Krallinger, M. and Valencia, A. (2005) Text-mining and information-retrieval services for molecular biology, *Genome Biol*, 6, 224.

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, 103-107.

Lakowicz, J.R. (1988) Principles of frequency-domain fluorescence spectroscopy and applications to cell membranes, *Subcell Biochem*, 13, 89-126.

Lee, S.A., Chan, C.H., Tsai, C.H., Lai, J.M., Wang, F.S., Kao, C.Y. and Huang, C.Y. (2008) Ortholog-based protein-protein interaction prediction and its application to inter-species interactions, *BMC Bioinformatics*, 9 Suppl 12, S11.

Leinonen, R., Nardone, F., Zhu, W. and Apweiler, R. (2006) UniSave: the UniProtKB sequence/annotation version database, *Bioinformatics*, 22, 1284-1285.

Li, L., Stoeckert, C.J., Jr. and Roos, D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes, *Genome Res*, 13, 2178-2189.

Li, S., Armstrong, C.M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.O., Han, J.D., Chesneau, A., Hao, T., Goldberg, D.S., Li, N., Martinez, M., Rual, J.F., Lamesch, P., Xu, L., Tewari, M., Wong, S.L., Zhang, L.V., Berriz, G.F., Jacotot, L., Vaglio, P., Reboul, J., Hirozane-Kishikawa, T., Li, Q., Gabel, H.W., Elewa, A., Baumgartner, B., Rose, D.J., Yu, H., Bosak, S., Sequerra, R., Fraser, A., Mango, S.E., Saxton, W.M., Strome, S., Van Den Heuvel, S., Piano, F., Vandenhaute, J., Sardet, C., Gerstein, M., Doucette-Stamm, L., Gunsalus, K.C., Harper, J.W.,

Cusick, M.E., Roth, F.P., Hill, D.E. and Vidal, M. (2004) A map of the interactome network of the metazoan C. elegans, *Science*, 303, 540-543.

Li, W.H., Yang, J. and Gu, X. (2005) Expression divergence between duplicate genes, *Trends Genet*, 21, 602-607.

Matsumura, H., Reuter, M., Kruger, D.H., Winter, P., Kahl, G. and Terauchi, R. (2008) SuperSAGE, *Methods Mol Biol*, 387, 55-70.

Modlin, R.L. and Doherty, P. (2003) Host defense against microbial pathogens – the immune system's weapons of mass destruction, *Current Opinion in Immunology*, 15, 393–395.

Moore, B. (2004) Bifunctional and moonlighting enzymes: lighting the way to regulatory control, *Trends Plant Sci*, 9, 221-228.

Ng, S.K., Zhang, Z. and Tan, S.H. (2003) Integrative approach for computationally inferring protein domain interactions, *Bioinformatics*, 19, 923-929.

O'Connell, M.R., Gamsjaeger, R. and Mackay, J.P. (2009) The structural analysis of protein-protein interactions by NMR spectroscopy, *Proteomics*, 9, 5224-5232.

Pearson, W.R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA, *Methods Enzymol*, 183, 63-98.

Periasamy, A. (2001) *Methods in Cellular imaging*. New York.

Persico, M., Ceol, A., Gavrila, C., Hoffmann, R., Florio, A. and Cesareni, G. (2005) HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms, *BMC Bioinformatics*, 6 Suppl 4, S21.

Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, *Nucleic Acids Res*, 35, D61-65.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R. and Lopez, R. (2005) InterProScan: protein domains identifier, *Nucleic Acids Res*, 33, W116-120.

Ramm, M., Dangoor, K. and Sayfan, G. (2007) *Rapid web applications with TurboGears*. R.R. Donelley and Sons Boston.

Remm, M., Storm, C.E. and Sonnhammer, E.L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons, *J Mol Biol*, 314, 1041-1052.

Rodland, K.D., Adkins, J.N., Ansong, C., Chowdhury, S., Manes, N.P., Shi, L., Yoon, H., Smith, R.D. and Heffron, F. (2008) Use of high-throughput mass spectrometry to elucidate host-pathogen interactions in Salmonella, *Future Microbiol*, 3, 625-634.

Rual, J.F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D.S., Zhang, L.V., Wong, S.L., Franklin, G., Li, S., Albala, J.S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R.S., Vandenhaute, J., Zoghbi, H.Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M.E., Hill, D.E., Roth, F.P. and Vidal, M. (2005) Towards a proteome-scale map of the human protein-protein interaction network, *Nature*, 437, 1173-1178.

Saha, S., Sparks, A.B., Rago, C., Akmaev, V., Wang, C.J., Vogelstein, B., Kinzler, K.W. and Velculescu, V.E. (2002) Using the transcriptome to annotate the genome, *Nat Biotechnol*, 20, 508-512.

Sayers, E.W., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L.Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D.J., Madden, T.L., Maglott, D.R., Miller, V., Mizrachi, I., Ostell, J., Pruitt, K.D., Schuler, G.D., Sequeira, E., Sherry, S.T., Shumway, M., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusova, T.A., Wagner, L., Yaschenko, E. and Ye, J. (2009) Database resources of the National Center for Biotechnology Information, *Nucleic Acids Res*, 37, D5-15.

Shalon, D., Smith, S.J. and Brown, P.O. (1996) A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization, *Genome Res*, 6, 639-645.

Shin, D.H., Hou, J., Chandonia, J.M., Das, D., Choi, I.G., Kim, R. and Kim, S.H. (2007) Structure-based inference of molecular functions of proteins of unknown function from Berkeley Structural Genomics Center, *J Struct Funct Genomics*, 8, 99-105.

Sjolander, K. (2004) Phylogenomic inference of protein molecular function: advances and challenges, *Bioinformatics*, 20, 170-179.

Smith, H. (1995) The revival of interest in mechanisms of bacterial pathogenicity, *Biol Rev Camb Philos Soc*, 70, 277-316.

Smith, T.F., Waterman, M.S. and Burks, C. (1985) The statistical distribution of nucleic acid similarities, *Nucleic Acids Res*, 13, 645-656.

Sprinzak, E. and Margalit, H. (2001) Correlated sequence-signatures as markers of protein-protein interaction, *J Mol Biol*, 311, 681-692.

Stellberger, T., Hauser, R., Baiker, A., Pothineni, V.R., Haas, J. and Uetz, P. Improving the yeast two-hybrid system with permutated fusions proteins: the Varicella Zoster Virus interactome, *Proteome Sci*, 8, 8.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksoz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H. and Wanker, E.E. (2005) A human protein-protein interaction network: a resource for annotating the proteome, *Cell*, 122, 957-968.

Szarka, E., Mityko, J., Szarka, J. and Csillery, G. (2002) Interaction between the generla and the specific plant defense reactions.

Uniprot [http://www.uniprot.org].

Van Dongen, S. (2000) Graph clustering by flow simulation. *Centre for Mathematics and Computer Science*. University of Utrecht, Netherlands.

van Ruissen, F. and Baas, F. (2007) Serial analysis of gene expression (SAGE), *Methods Mol Biol*, 383, 41-66.

Velculescu, V.E., Vogelstein, B. and Kinzler, K.W. (2000) Analysing uncharted transcriptomes with SAGE, *Trends Genet*, 16, 423-425.

Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995) Serial analysis of gene expression, *Science*, 270, 484-487.

Verschure, P.J., Visser, A.E. and Rots, M.G. (2006) Step out of the groove: epigenetic gene control systems and engineered transcription factors, *Adv Genet*, 56, 163-204.

Vignali, M., McKinlay, A., LaCount, D.J., Chettier, R., Bell, R., Sahasrabudhe, S., Hughes, R.E. and Fields, S. (2008) Interaction of an atypical Plasmodium falciparum ETRAMP with human apolipoproteins, *Malar J*, 7, 211.

Walker, D.H., Valbuena, G.A. and Olano, J.P. (2003) Pathogenic mechanisms of diseases caused by Rickettsia, *Ann N Y Acad Sci*, 990, 1-11.

Walliker, D., Quakyi, I.A., Wellems, T.E., McCutchan, T.F., Szarfman, A., London, W.T., Corcoran, L.M., Burkot, T.R. and Carter, R. (1987) Genetic analysis of the human malaria parasite Plasmodium falciparum, *Science*, 236, 1661-1666.

Wang, X. and Feuerstein, G.Z. (1997) The use of mRNA differential display for discovery of novel therapeutic targets in cardiovascular disease, *Cardiovasc Res*, 35, 414-421.

Wellems, T.E., Panton, L.J., Gluzman, I.Y., do Rosario, V.E., Gwadz, R.W., Walker-Jonah, A. and Krogstad, D.J. (1990) Chloroquine resistance not linked to mdr-like genes in a Plasmodium falciparum cross, *Nature*, 345, 253-255.

Winnenburg, R., Baldwin, T.K., Urban, M., Rawlings, C., Kohler, J. and Hammond-Kosack, K.E. (2006) PHI-base: a new database for pathogen host interactions, *Nucleic Acids Res*, 34, D459-464.

Xenarios, I., Salwinski, L., Duan, X.J., Higney, P., Kim, S.M. and Eisenberg, D. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions, *Nucleic Acids Res*, 30, 303-305.

Xiang, Z., Tian, Y. and He, Y. (2007) PHIDIAS: a pathogen-host interaction data integration and analysis system, *Genome Biol*, 8, R150.

Yu, H., Luscombe, N.M., Lu, H.X., Zhu, X., Xia, Y., Han, J.D., Bertin, N., Chung, S., Vidal, M. and Gerstein, M. (2004) Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs, *Genome Res*, 14, 1107-1118.

Zhou, D. and He, Y. (2008) Extracting interactions between proteins from the literature, *J Biomed Inform*, 41, 393-407.

Zinsser, H. (1914) Infection and the problem of virulence. In, *Infection and resistance*. The Macmillan Company, New York, 1-27.