In silico drug discovery on computational Grids for finding novel drugs against neglected diseases

Dissertation zur
Erlangung des Doktorgrades (Dr. rer. nat.) der
Mathematisch-Naturwissenschaftlichen Fakultat der
Rheinischen Friedrich-Wilhelms-Universitat Bonn
vorgelegt von

Vinod Kumar Kasam
Aus Warangal, Indien

Bonn

September 2009

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn.

1. Referent: Univ.-Prof. Dr. Martin Hofmann-Apitius
2. Referent: Univ.-Prof. Dr. Christa Mueller

Tag der Promotion: 30.04.2010

Erscheinungsjahr: 2010

For my Family: My Wife and Son

Abstract

Malaria is a dreadful disease affecting 300 million people and killing 1-1.5 million people every year. Malaria is caused by a protozoan parasite, belonging to the genus Plasmodium. There are several species of Plasmodium infecting cattle, birds, and humans. The four species *P.falciparum*, *P.vivax*, *P.malariae and P.ovale* are in particular considered important, as these species infect humans. One of the main causes for the comeback of malaria is that the most widely used drug against malaria, chloroquine, has been rendered useless by drug resistance in much of the world. New antimalarial drugs are presently available but the potential emergence of resistance, the difficulty to synthesize these drugs at a large-scale and their cost make it of utmost importance to keep searching for new drugs.

Despite continuous efforts of the international community to reduce the impact of malaria on developing countries, no significant progress has been made in the recent years and the discovery of new drugs is more than ever needed. Out of the many proteins involved in the metabolic activities of the Plasmodium parasite, some are promising targets to carry out rational drug discovery.

In silico drug design, especially vHTS is a widely and well-accepted technology in lead identification and lead optimization. This approach, therefore builds upon the progress made in computational chemistry to achieve more accurate *in silico* docking and in information technology to design and operate large-scale Grid infrastructures. One potential limitation of structure-based methods, such as molecular docking and molecular dynamics is that; both are computational intensive tasks. Recent years have witnessed the emergence of Grids, which are highly distributed computing infrastructures particularly well fitted for embarrassingly parallel computations such as docking and molecular dynamics.

The current thesis is a part of WISDOM project, which stands for Wide *In silico* Docking on Malaria. This thesis describes the rational drug discovery activity at large-scale, especially molecular docking and molecular dynamics on computational Grids in finding hits against four different targets (PfPlasmepsin, PfGST, PfDHFR, PvDHFR (wild type and mutant forms) implicated in malaria.

The first attempt at using Grids for large-scale virtual screening (combination of molecular docking and molecular dynamics) focused on plasmepsins and ended up in the identification of previously unknown scaffolds, which were confirmed *in vitro* to be active plasmepsin inhibitors. The combination of docking and molecular dynamics simulations, followed by rescoring using sophisticated scoring functions resulted in the identification of 26 novel sub-micromolar inhibitors. The inhibitors are further clustered into five different scaffolds. While two scaffolds, diphenyl urea, and thiourea analogues are already known as plasmepsin inhibitors, albeit the compounds identified here are different from the existing ones, with the new class of potential inhibitors, the guanidino group of compounds, we have established a new class of chemical entities with inhibitory activity against *Plasmodium falciparum* plasmepsins.

Following the success achieved on plasmepsin, a second drug finding effort was performed, focussed on one well known target, dihydrofolate reductase (DHFR), and on a new promising one, glutathione-S-transferase. Modeling results are very promising and based on these *in silico* results, in vitro tests are in progress.

Thus, with the work presented here, we not only demonstrate the relevance of computational grids in drug discovery, but also identify several promising small molecules (success achieved on *P. falciparum* plasmepsins). With the use of the EGEE infrastructure for the virtual screening campaign against the malaria-causing parasite *P. falciparum*, we have demonstrated that resource sharing on an e-Science infrastructure such as EGEE provides a new model for doing collaborative research to fight diseases of the poor.

Through WISDOM project, we propose a Grid-enabled virtual screening approach, to produce focus compound libraries for other biological targets relevant to fight the infectious diseases of the developing world.

Acknowledgements

I am grateful to numerous local and global persons who have contributed towards my thesis. Firstly, I thank Prof. Dr. Martin Hofmann-Apitius for giving me an opportunity to do my PhD thesis at Fraunhofer-SCAI, Germany. His encouragement always motivated me to focus beyond my work. As my supervisor, he has constantly motivated me to remain focused on achieving my goal. I am thankful to Prof. Dr. Christa Mueller for her readiness to be co-supervisor on the thesis.

I am very grateful to Dr. Vincent Breton, LPC, IN2P3-CNRS, Clermont-Ferrand France for his guidance, support and providing me a chance to work in his lab, without which this thesis would have not been possible.

I want to thank Prof. Giulio Rasteli, University of Modena, Italy for his guidance and training on the molecular dynamics approach. I thank Prof. Doman Kim, University of South Korea, for kindly performing the in vitro tests. At the outset, I would like to express my special thanks and regards to Jean Salzemann, Marc Zimmermann, Astrid Maass, Antje Wolf and Mohammed Shahid for their help and scientific discussions.

My special thanks to Ana Da Costa and Nicolas Jacq. I sincerely feel that working together with them was beneficial for my successful completion of the thesis.

I thank all my colleagues at Fraunhofer-SCAI and LPC, IN2P3-CNRS for their immense support and co-operation during my thesis work.

My very special thanks to all the people involved in WISDOM collaboration.

List of Abbreviations

Plm	Plasmespin
MD	Molecular Dynamics
MOE	Molecular Operating Environment
vHTS	Virtual High Throughput Screening
HTS	High Throughput Screening
DHFR	Dihydrofolate Reductase
RMSD	Root Mean Square Deviation
EGEE	European Grid Enabling E-science
GST	Glutathione-S-Trasferase
MM-PBSA	Molecular Mechanics Poisson Boltzmann Surface Area
MM-GBSA	Molecular Mechanics Generalized Born Surface Area
NCE	New Chemical Entity
ADME	Absorption, Distribution, Metabolism, Elimination

Contents

List of Figures

List of Tables

List of Publications

PATENT

1. Doman Kim, Hee Kyoung Kang, Do Won Kim, Giulio Rastelli, Ana-Lucia Da Costa, **Vinod Kasam,** Vincent Breton. "*Pharmaceutical composition for preventing and treating malaria comprising compounds that inhibit Plasmepsin II activity and the method of treating malaria using thereof*". Priority number KR 20080037148 20080422

Publications

2. **Vinod Kasam,** Jean Salzemann, Marli Botha, Ana Dacosta, Gianluca Degliesposti, Raul Isea, Doman Kim, Astrid Maass, Colin Kenyon, Giulio Rastelli, Martin Hofmann-Apitius, Vincent Breton. WISDOM-II: Screening against multiple targets implicated in malaria using computational grid infrastructures. **Malaria Journal, 2009, 8:88. [HIGHLY ACCESSED]**

3. **Vinod Kasam.,** Zimmermann, M., Maaß, A., Schwichtenberg, H., Wolf, A., Jacq, N., Breton, V., Hofmann, M. Design of Plasmepsin Inhibitors: A Virtual High Throughput Screening Approach On The EGEE Grid, **J. Chem. Inf. Model. 2007, 47, 1818-1828**

4. **Vinod Kasam,** Jean Salzemann, Nicolas Jacq, Astrid Mass and Vincent Breton. Large-scale Deployment of Molecular Docking Application on Computational Grid infrastructures for Combating Malaria. ccgrid, pp. 691-700, **Seventh IEEE International Symposium on Cluster Computing and the Grid (CCGrid '07), 2007.**

5. Degliesposti G, **Vinod Kasam**, Da Costa A, Kim D, Hee-Kyoung K, Do-Won Kim, Breton V, Rastelli G. Design and Discovery of novel plamepsin inhibitors using automated work flow on large-scale grids. **ChemMedChem 2009, 4(7):1164-73.**

6. Younesi E, **Kasam V,** Hofmann-Apitius M. Direct Use of Information Extraction from Scientific Text for Modeling and Simulation in the Life Sciences. **Journal Library Hi Tech. 2009, 27(4), 505-519.**

7. Wolf A, Hofmann-Apitius M, Moustafa G, Azam N, Kalaitzopolous D, Yu K, **Kasam V**. Dock flow – A prototypic pharma grid for virtual screening integrating four different docking tools. **Stud Health Technol Inform. 2009, 147:3-12.**

8. Wolf, A, Shahid, M., **Kasam V,** Hofmann-Apitius, M. In silico drug discovery approaches on grid computing infrastructures. **Current Clinical Pharmacology, 2010, 5, 37-46.**

9. Birkholtz, L.-M., Bastien, O., Wells, G., Grando, D., Joubert, F., **Kasam, V.,** Zimmermann, M., Ortet, P., Jacq, N., Saidani, N., Hofmann-Apitius, S., Hofmann-Apitius, M., Breton, V., Louw, A.I., Marechal, E. Integration and mining of malaria molecular, functional and pharmacological data: how far are we from a chemogenomic knowledge space?, **Malar J. 2006; 5: 110** **[HIGHLY ACCESSED]**

10. Jacq, N., Salzemann, J., Jacq, F., Legré, Y., Medernach, E., Montagnat, J., Maaß, A., Reichstadt, M., Schwichtenberg, H., Sridhar, M., **Kasam, V.,** Zimmermann, M., Hofmann, M., Breton, V. Grid-enabled Virtual Screening against malaria. **J. Grid Comput. 6(1): 29-43 (2008).**

11. Jacq, N., Breton, V., Chen, H.-Y., Ho, L.-Y., Hofmann, M., Lee, H.-C., Legré, Y., Lin, S.C., Maaß, A., Medernach, E., Merelli, I., Milanesi, L., Rastelli, G., Reichstadt, M., Salzemann, J., Schwichtenberg, H., Sridhar, M., **Kasam, V.,** Wu, Y.-T., Zimmermann, M., Virtual Screening on Large-scale Grids**. Parallel Computing 33(4-5): 289-301 (2007).**

12. Shahid. M, Ziegler. W, **Kasam.V,** Zimmermann. M, Hofmann-Apitius. M. Virtual High Throughput Screening on Optical High Speed Network. **Stud Health Technol Inform. 2008; 138: 124-34.**

13. Breton, V., Jacq, N**., Kasam, V.,** Hofmann-apitius, M., Grid Added Value to Address Malaria. **IEEE Trans Inf Technol Biomed. 2008 Mar;12(2):173-81**

14. Robbie P. Joosten, Jean Salzemann, Christophe Blanchet, Vincent Bloch, Vincent Breton, Ana L. Da Costa, **Vinod Kasam,** Vincent Breton, Gert Vriend et al. Re-refinement of all X-ray structures in the PDB. **J. Appl. Cryst. (2009), 42, 1-9.**

15. Breton, V., Jacq, N., **Kasam, V.,** Salzemann, J., Chapter 9: Deployment of Grid life sciences applications. Talbi, E.-G., Zomaya, A. (eds.) **Grids for Bioinformatics and Computational Biology. Wiley-Interscience 2007.**

1 Chapter1. Introduction

Diseases affecting the poor are widely ignored by the pharmaceutical industry. They are known as neglected diseases. These diseases are often caused by parasites, worms and bacteria. The parasitic and bacterial infections include three soil-transmitted helminth infections (ascariasis, hookworm infection, and trichuriasis), lymphatic filariasis, onchocerciasis, dracunculiasis, schistosomiasis, Chagas disease, human African trypanosomiasis, leishmaniasis, Buruli ulcer, leprosy, trachoma, treponematoses, leptospirosis, strongyloidiasis, foodborne trematodiases, neurocysticercosis, scabies and infectious parasitic diseases including diseases such as Malaria, Dengue fever, Kalaazar, Toxoplosmosis. Table 1 describes some of the neglected diseases and their respective causative organisms [1, 2].

Disease	Organism	Scope	Therapy needs
Malaria	*Plasmodium* spp.	500 million infections annually	Novel drugs and Circumventing drug resistance
Leishmaniasis	*Leishmania* spp	2 million infections annually	Safe, orally bioavailable drugs, especially for the visceral form of the disease
Trypanosomiasis (sleeping sickness, Chagas disease)	*T. brucei* (sleeping sickness) *T. cruzi* (Chagas disease)	HAT: 300,000 cases annually Chagas: 16 million existing infections	Safe, orally bioavailable drugs, especially for the chronic phases of disease
Schistosomiasis	*Schistosoma* spp.	>200 million existing infections	Backup drug should resistance arise to praziquantel
Giardiasis/amebiasis *Giardia lamblia*;	*Entamoeba histolytica*	Millions of cases of diarrhea annually	Well-tolerated drugs
Ascariasis	*Ascaris lumbricoides*	807 Millions	Access to essential medicines
Leprosy	*Mycobacterium leprae*	0.4 millions	Access to essential medicines
Hookworm infection	*Ancylostoma duodenale*	576 millions	Access to essential medicines and high efficacy
Lymphatic filariasis	*Wuchereria bancrofti,*	120 millions	Access to essential medicines
Trachoma	*Chlamydia trachomatis*	84 millions	Access to essential medicines and needs public health interventions

Table 1: Demonstrates the spread of neglected diseases, adapted from [1, 2]
The Table illustrates some of the most worst tropical diseases of the world, organism responsible for the disease, scope of the disease and therapy needs.

Status on drug discovery related against neglected diseases

More than $100 billion is spent per year on health research and drug development by pharmaceutical industries and other sources, but less than 10 percent is spent on 90 percent of the world's health problems affecting the poor of Africa, Asia, and Latin America. There is an urgent need to correct the fatal imbalance of the current drug development model, which is currently accepting a death toll of 14 million people from infectious diseases each year. At present, the majority of medicines are being developed by rich nations whose inhabitants can afford expensive and often complicated drug therapies that are either too costly or too complicated or both for nations struggling against poverty and disease epidemics [3].

As most patients with such diseases live in developing countries and are too poor to pay for expensive drugs, the pharmaceutical industry has traditionally ignored these diseases. Over the past decade, however, the public sector, by creating favorable marketing conditions, has persuaded industry to enter into public private partnerships to tackle neglected diseases such as malaria, HIV, and tuberculosis. This industry invests almost exclusively in developing drugs that are likely to be marketable and profitable drugs for conditions such as pain, cancer, heart disease, and baldness. Figure 1 and 2 illustrates the current state-of-the-art on diseases. Public policies, such as tax incentives and patent protection are geared towards this market driven private investment. As a result, out of 1393 new drugs marketed between 1975 and 1999, only 16 were for neglected diseases, yet these diseases accounted for over 10% of the global disease burden (Figure 1). In contrast, over two thirds of new drugs were "me too drugs" (modified versions of existing drugs), which do little or nothing to change the disease burden [4, 5]. The current thesis details about malaria in particular and describes the *in silico* drug discovery activities against potential malarial targets.

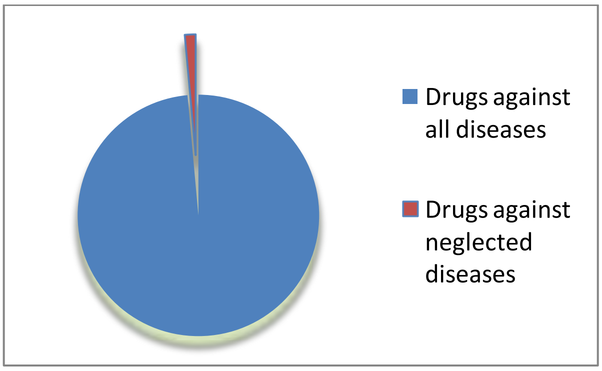


Figure 1: Number of drugs developed against neglected diseases over the years [4, 5]
This Figure gives the current state-of-the-art of drugs developed until 1999. It clearly demonstrates that very few drugs were developed for neglected diseases.

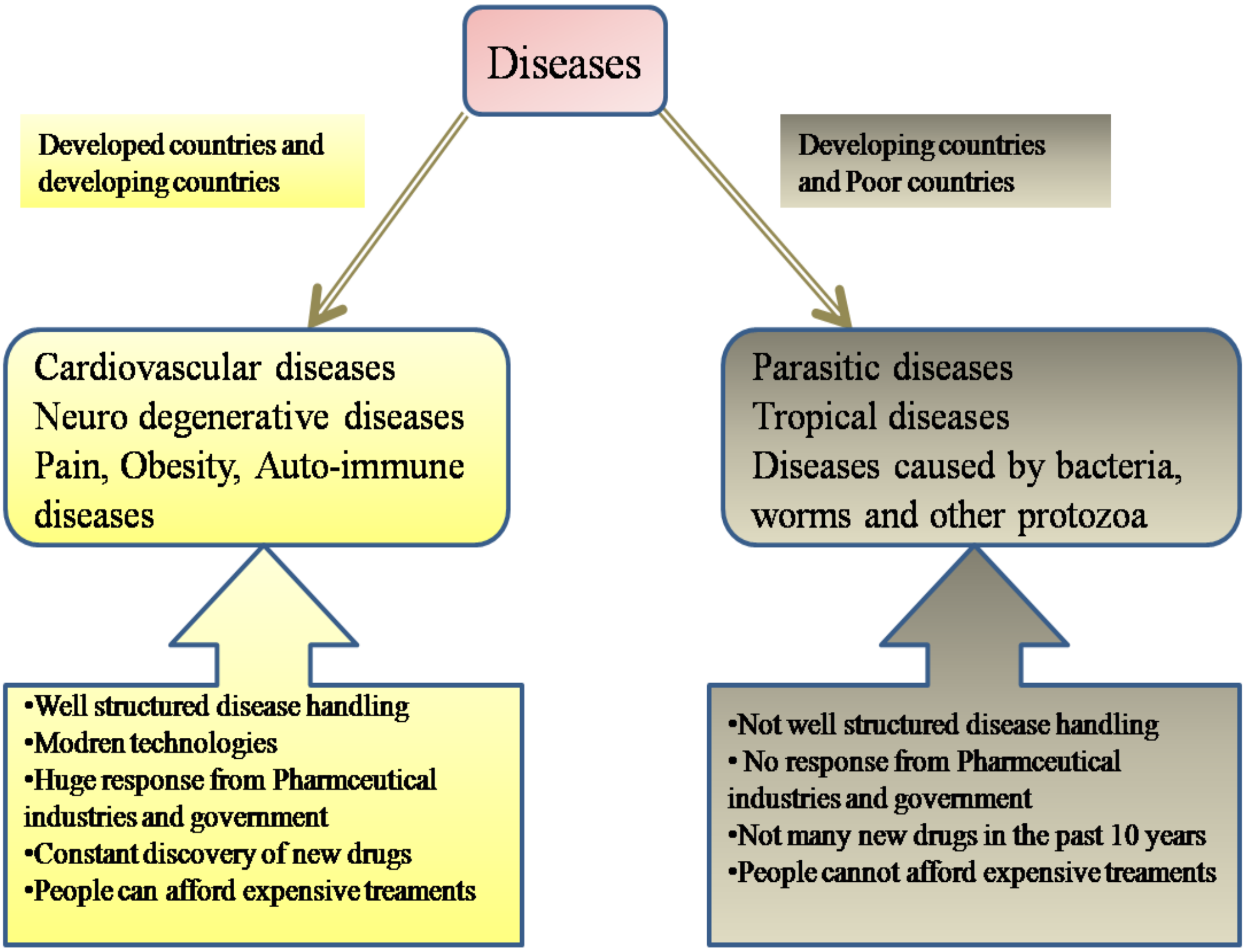

Diseases
Developed countries and developing countries
Developing countries and Poor countries
Cardiovascular diseases
Neuro degenerative diseases
Pain, Obesity, Auto-immune diseases
Parasitic diseases
Tropical diseases
Diseases caused by bacteria, worms and other protozoa
•Well structured disease handling
•Modren technologies
•Huge response from Pharmceutical industries and government
•Constant discovery of new drugs
•People can afford expensive treaments
•Not well structured disease handling
• No response from Pharmceutical industries and government
•Not many new drugs in the past 10 years
•People cannot afford expensive treaments

Figure 2 : Schematic representation of state-of-art-the of neglected diseases.
The Figure demonstrates that diseases have been segmented into neglected diseases and chronic diseases based on the diseases affected to people of developed nations and poor nations. It illustrates that neglected diseases are not handled well because of lack of pharmaceutical interest, and further because people living in these countries are poor to pay expensive treatments.

1.1 Malaria

Malaria is an infectious disease caused by the parasite called Plasmodium and is a serious problem for human health, especially to the so-called "Third World." There are four identified species of this parasite causing human malaria, namely, *Plasmodium vivax*, *P. falciparum*, P. ovale and P. malariae. The female anopheles mosquito transmits plasmodium species. It is a disease that can be treated in just 48 hours, yet it can cause fatal complications if the diagnosis and treatment are delayed. More than 2400 million people, over 40% of the world's population are affected by this disease in more than 100 countries in the tropics from South America to the Indian peninsula [6]. The tropics provide ideal breeding and living conditions for the anopheles mosquito, and hence this distribution. According to WHO, there were an estimated 247 million malaria cases among 3.3 billion people at risk in 2006, causing nearly a million deaths, mostly of children under 5 years. 109 countries were endemic for malaria in 2008, 45 of them within the WHO African region [7]. The geographical distribution of malaria, according to center for disease control in 2006 is shown in Figure 3. Every year 300 million to 500 million people suffer from this disease (90% of them in sub-Saharan Africa, two thirds of the remaining cases occur in six countries like India, Brazil, Sri Lanka, Vietnam, Colombia and Solomon Islands). WHO forecasts a 16% growth in malaria cases annually. About 1.5 million to 3 million people die of malaria every year (85% of these occur in Africa), accounting for about 45% of all fatalities in the world [8]. One child dies of malaria in Africa every 20 sec., and there is one malarial death every 12 sec somewhere in the world. Malaria kills in 1 year what AIDS killed in 15 years. In 15 years, if 5 million have died of AIDS, 50 million have died of malaria [9, 10].

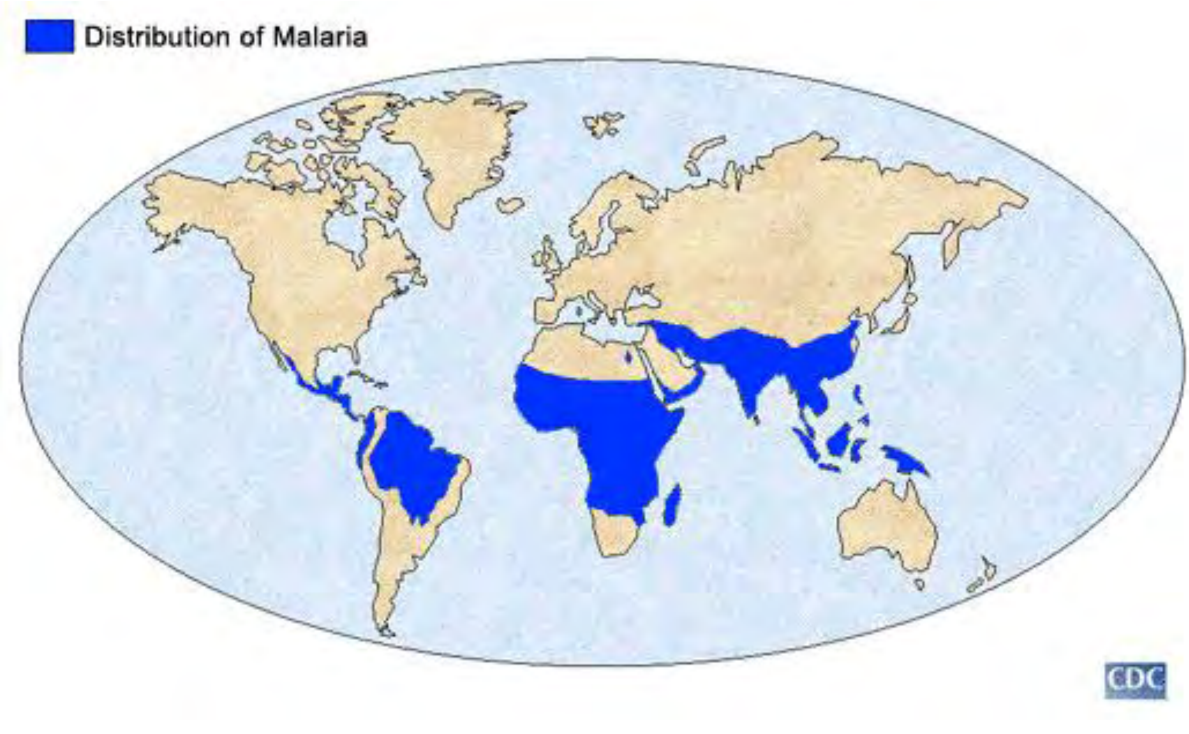


Figure 3 Spread of malaria all over the world by 2006 [8]
The Figure clearly illustrates that malaria is widely spread in Asia, Africa and to some countries in South America (Developing and underdeveloped countries). Courtesy: Center for Disease Control. Source: Wikipedia commons.

1.1.1 Complex life cycle of malaria

The first step for developing novel drugs against any disease is, understanding the disease. This section gives insight into the life cycle of malaria and its associated complexity.

Plasmodium complete life cycle involves both human (host) and female anopheles mosquito (insect vector). Figure 4 demonstrates the complete life cycle of plasmodium [11].

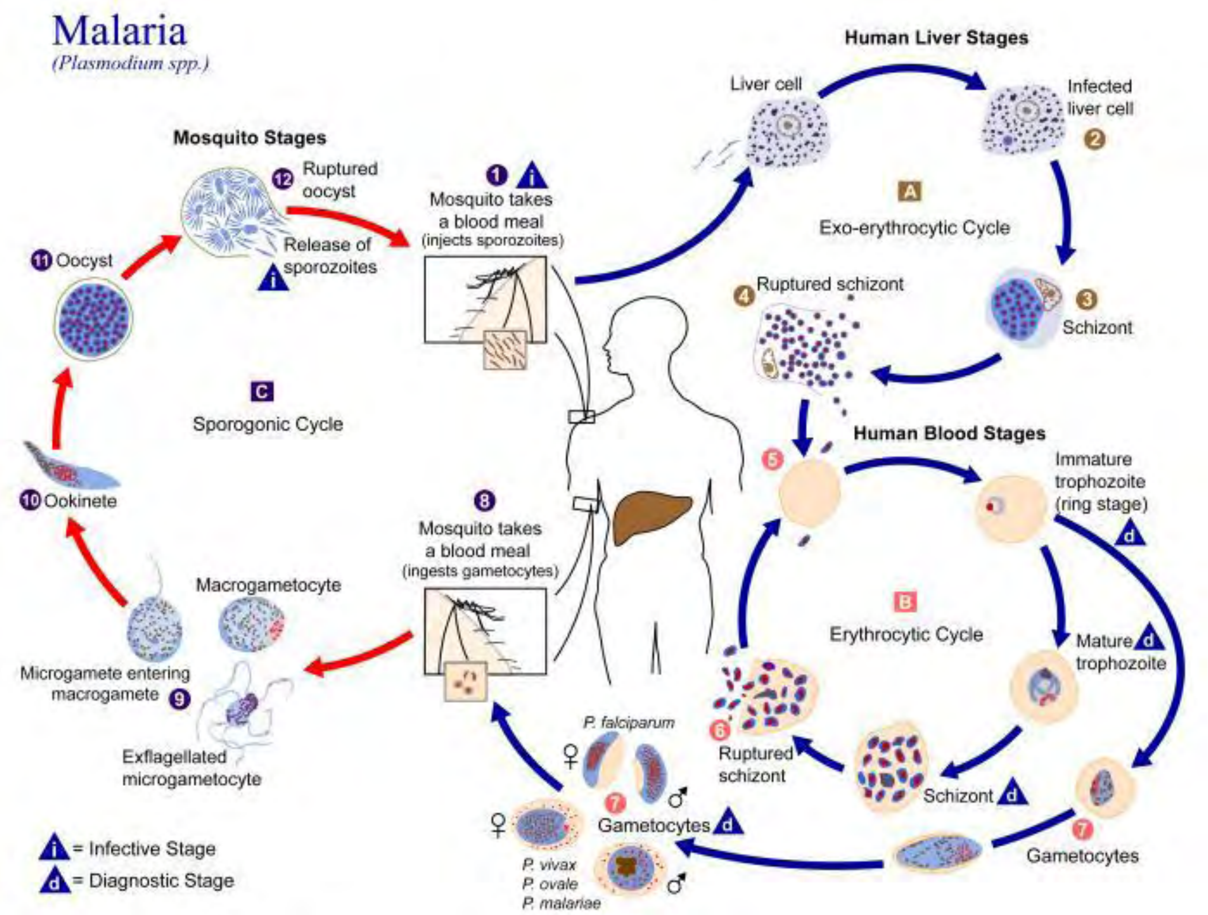


Figure 4: Complete life cycle of malaria causing Plasmodium species.
The Figure illustrates three different cycles that occur in human and mosquito. Different cycles are termed as A, B, C and numbers illustrates the various parasitic stages.
Courtesy: Center for Disease Control and preventions. Source: Wikipedia commons

As shown in the Figure 4, the life cycle of plasmodium is divided into three cycles,

A. Exo-erythrocytic cycle
B. Erythrocytic cycle
C. Sporogonic cycle

In each phase, plasmodium occurs in different forms and stages. These various stages of plasmodium help in the diagnosis of the disease and as well as in treating the disease. There are several drugs available in the market that can counteract a particular stage or several stages of the Plasmodium. Table 2 illustrates the currently available drugs inhibiting explicit stages and/or part of a cycle of Plasmodium life cycle.

Drug Class	Drugs	Stages of Plasmodium
8- Amino Quinolines	Primaquine, Tafenoquine	Hypnozoites, Gametocytes
4- Amino Quinolines	Chloroquine, Amidoquine	Intra-erythrocytic stages, Gametocytes
Quinoline-alcohols	Quinine, Mefloquine	Erythrocytic stages
Aryl-alcohols	Halofrantine, Pyronaridine	Erythrocytic stages
Antifolates	Proguanil, Pyrimethamine, Sulfadoxine, Dapsone	Erythrocytic stages
Artemesinins	Dihydroartemesinin, Artesunate, Artemether, Arteether,	Gametocytes
Antibiotics	Tetracyclin, Doxycycline,	Intra-erythrocytic stages

Table 2: Illustrates examples of currently available different classes of anti malarial drugs that are active against various stages of the plasmodium.

A. Exo-erythrocytic cycle

- In the Figure 4, the cycle A represents the exo-erythrocytic cycle. The exo-erythrocytic cycle is defined as the process occurring outside the erythrocytes (Exo= Outside and erythrocytes= red blood cells) in human. When a female anopheles mosquito carrying sporozoites feeds on the human, during this meal, the sporozoits are injected into the blood stream and later enters the liver and invades liver cells. Inside the hepatocytes the sporozoite develops into the trophozoite, where it undergoes several divisions and forming several schizonts. The schizont encapsulates membrane around itself and forms several merozoites. Some malaria parasite species remain dormant for extended periods in the liver, causing relapses weeks or months later [12, 8].

Erythrocytic cycle

- In the Figure 4, the cycle B represents the erythrocytic cycle. The erythrocytic cycle takes place inside the human red blood cells. The merozoites invade erythrocytes and undergo a trophic period in which the parasite enlarges. The early trophozoite is often referred to as 'ring form' because of its morphology. Trophozoite enlargement is accompanied by an active metabolism including the ingestion of host cytoplasm and the proteolysis of hemoglobin into amino acids. Plasmepsin, the target protein of the current study is an aspartic protease initiates the hemoglobin degradation. More details about hemoglobin degradation and the role plasmepsin family of proteins are given in chapter 4. Some of the merozoite-infected blood cells leave the cycle of asexual multiplication. Instead of replicating, the merozoites in these cells develop into sexual forms of the parasite, called male and female gametocytes, which circulate in the bloodstream [13, 10, 8].

Sporogonic cycle

- In the Figure 4, the cycle C represents the Exo-erythrocytic cycle. When a mosquito bites an infected human, it ingests the gametocytes. In the mosquito gut, the infected human blood cells burst, releasing the gametocytes, which develop further into mature sex cells called gametes. Male and female gametes fuse to form diploid (cells containing full set of chromosomes) zygotes, which develop into actively moving ookinetes that burrow into the mosquito midgut wall and form oocysts.
- Growth and division of each oocyst produces thousands of active haploid forms called sporozoites. After 8-15 days (depending upon the plasmodium species), the oocyst bursts, releasing sporozoites into the body cavity of the mosquito, from which they travel to and invade the mosquito salivary glands. The cycle of human infection restarts when the mosquito takes a blood meal, injecting the sporozoites from its salivary glands into the human blood stream [13, 10, 8].

1.1.2 Current drugs

There are several antimalarial drugs presently available. In most cases, antimalarial drugs are targeted against the asexual erythrocytic stage of the parasite. The parasite degrades hemoglobin in its acidic food vacuole, producing free heme able to react with molecular oxygen and thus to generate reactive oxygen species as toxic by-products. A major pathway of detoxification of heme moieties is polymerization as malaria pigment [14, 15]. The majority of antimalarial drugs act by disturbing the polymerization (and/or the detoxification by any other way) of heme, thus killing the parasite with its own metabolic waste.

The most widely used are quinine and its derivatives and antifolate combination drugs. The main classes of active schizontocides are 4-aminoquinolines, aryl-alcohols including quinoline-alcohols and antifolate compounds which inhibit the synthesis of parasitic pyrimidines. The newest class of antimalarials is based on the natural endoperoxide artemisinin and its hemisynthetic derivatives and synthetic analogs. Some antibiotics are also used, generally in association with quinoline-alcohols [16, 17]. Few compounds are active against gametocytes and also against the intra-hepatic stages of the parasite [18].

Artemisinin compounds

A number of sesquiterpine lactone compounds have been synthesized from the plant *Artemisia annua* (artesunate, artemether, arteether) [18]. These compounds are used for treatment of severe malaria; furthermore, these compounds have shown very rapid parasite clearance times and faster fever resolution than that occurs with quinine. In some areas of South-East Asia, combinations of artemisinins and mefloquine offer the only reliable treatment for even uncomplicated malaria, due to the development and prevalence of multidrug resistant *P. falciparum* malaria [19, 20]. Combination therapy (an artemisinin compound given in combination with another antimalarial, typically a long half-life drug like mefloquine) has reportedly been responsible for inhibiting intensification of drug resistance and for decreased malaria transmission levels in South-East Asia [19, 21].

Challenges

Despite the availability of effective antimalarial drugs, which are capable of inhibiting various stages of the parasite, treatment of malaria is still with many challenges and limitations. Major challenges include:

a. Lack of epidemiological data and exact numbers of people dying due to illness in endemic countries.
b. Poor mosquito control, due to resistance of anopheles mosquito to the insecticides such as DDT.
c. Poor diagnosis
d. Unavailability of vaccination.
e. Delivering the drugs to the patients in need of the drugs.
f. Effective combination therapies that are frontline treatments are too expensive to be paid by the patients.
g. No new drugs in the past years, and resistance to existing malarial drugs.
h. Resistance to existing malarial drugs.

Drug resistance is principal challenge in tackling malaria; hence, it is further discussed in detail.

Drug resistance

According to Bruce-Chwatt LJ [22, 23], antimalarial drug resistance has been defined as the "ability of a parasite strain to survive and/or multiply despite the administration and absorption of a drug given in doses equal to or higher than those usually recommended but within tolerance of the subject". This definition was later modified to specify that the drug in question must "gain access to the parasite or the infected red blood cell for the duration of the time necessary for its normal action" [23].

Drug resistance has emerged towards all classes of antimalarials except for the artimisinins [24]. There is a threat of even resistance to artimisinin derivatives, as it has been already observed in the murine P. yoelii parasite [25]. Resistance of *P. falciparum* to chloroquine, the cheapest and the most commonly used drug is spreading in almost all the endemic countries. Resistance to the combination of sulfadoxine-pyrimethamine, which was already present in South America and in South-East Asia, is now emerging in East Africa also [10].

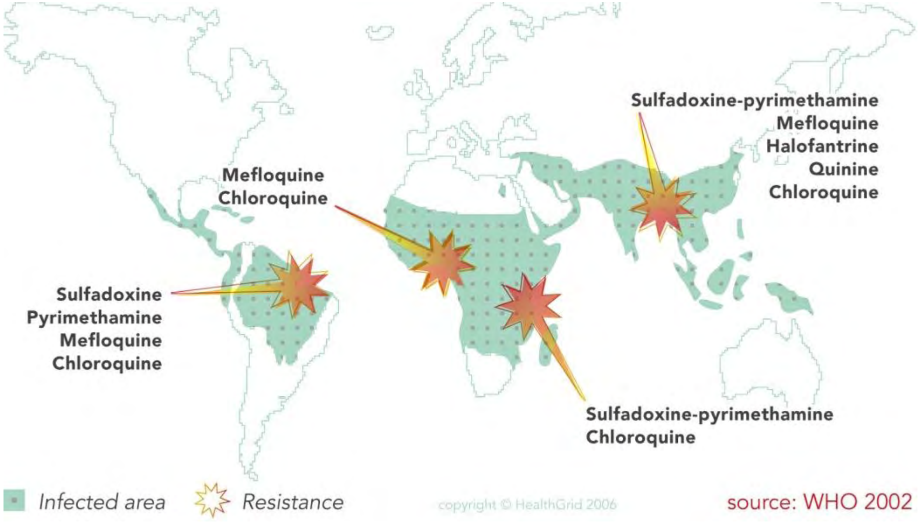


Figure 5 Geographical distribution of resistance to existing drugs of malaria [10]
This Figure illustrates that drug resistance is emerged for most of the existing anti-malarial and even combination therapies.

The molecular mechanisms behind the resistance depend on the chemical class of the compound and its mechanism of action. According to Peter B. Bloland [10], generally resistance appears to occur through spontaneous mutations that confer reduced sensitivity to a given drug or class of drugs. For some drugs, only a single point mutation is required to confer resistance, while for other drugs, multiple mutations appear to be required. When the mutations are not deleterious to the existence or reproduction of the parasite, drugs will eliminate the susceptible parasites while resistant parasites stay alive. Single malaria isolates have been found to be made up of heterogeneous populations of parasites that can have widely varying drug response characteristics, from highly resistant to completely sensitive [26]. Similarly, within a geographical area, malaria infections demonstrate a range of drug susceptibility. Over time, resistance will be established in the population and can be very stable; persisting long after specific drug pressure is removed. Geographical distribution of resistance to existing drugs worldwide is displayed in Figure 5.

Resistance to any new therapeutic agents is expected. Strategies to lengthen the drug lifetime are combination drug therapy and use of old drugs, wherever they are still effective [27].

Current International efforts in combating the disease

Most of the international efforts to counter malaria and other neglected diseases are philanthropic and public-private partnerships (PPP) [2, 28, 29]. PPP is a comprehensive framework, which aims at providing preventive chemotherapy packages, and further aims at developing, testing, and distributing a new generation tools to control these neglected diseases [1]. Generally, the private sector includes pharmaceutical companies, where they look for profit and the non-profit sector includes charities, foundations, and philanthropic institutions groups. The public sector includes international organizations, development and aid agencies, governments, and academia. Mefloquine, a potent antimalarial drug was discovered by WRAIR (US Walter Reed Army Institute of Research) [30] and was later developed by TDR (Tropical Disease Research) and the pharmaceutical industry. This collaborative effort between TDR [31] and WRAIR is a typical example of success achieved by PPP [4]. There were various examples of such collaborative efforts during 1990’s for antimalarial drug development. However, due to limited return on investment, there has been constant withdrawal of pharmaceutical industries from developing drugs against malaria. Due to this, the gap widened between the discovery stage and development process and thus halted the discovery of new chemical entities (NCE). To address this problem, there were some agreements between the public and private partners based on their coincidence of priorities of private and public sectors and thus both the public and private sectors contribute funds to develop a specific product. The collaboration between TDR, the Japanese government, and the Japanese pharmaceutical industry is one example of such partnerships [4]. World Health Organization (WHO) [10], Drugs for Neglected Diseases initiative DNDi [32], TDR [31], Malaria Vaccine Initiative (Grant of the Bill and Melinda Gates Foundation) [33], Medicines for Malaria Venture (MMV) [34], Roll Back Malaria initiative which was announced by WHO [35], Wellcome Trust [36], Sandler Family Supporting Foundation [2], St. Jude

Research Foundation [37] are some of the public organizations that are leading most of the efforts related to antimalarial drug development [2, 38].

Of the above mentioned, MMV is particularly interesting, because it is malaria based non-profit initiative and accounts for most of current antimalarial drug development projects. Another similar organization is the “Global Alliance for Tuberculosis”. It aims at development of drugs against tuberculosis. The aim of MMV is to convert the drug candidates into registered entities based on a social venture capital model funded by PPP. They rely on business-drug-development model, and derive short term funding requirements from the Melinda Gates foundation. Overall, MMV manages a portfolio of 21 projects that are at different stages of drug research and development process. Interestingly these projects target not a particular target or particular family of proteins but they target various enzymes and proteins belonging to various pathways that are essential for parasite survival. In 2008, MMV along with its industrial partnerships such as Glaxo SmithKline and Novartis enabled two molecules (Artemesinin Combination Treatments) to enter phase-I clinical trials and are for the first time tested in humans, besides that MMV enhanced four other chemical entities to enter into pre-clinical studies. MMV hopes to register its first drug in 2010 [38].

The Global Fund to Fight AIDS Tuberculosis and Malaria (GFTAM) is another active organization, which was established in January 2002 as an independent financing body to attract, manage, and disburse funds to AIDS, Tuberculosis, and Malaria [39].

1.1.3 Motivation

Despite continuous efforts of the international community to reduce the impact of malaria on poor and developing countries, there is steadily rise in the number of malarial infections and no significant progress in finding new drugs has been made in the recent years. Adding to the worse, currently available antimalarial drugs are losing effectiveness due to the emergence and spread of resistant parasite strains. In order to regain control over the disease, new drugs and treatments are urgently needed. Drug discovery efforts in this direction are most likely to be successful if they target a novel mechanism of action. Such approaches will lead to anti-malarial medicines that are functionally and structurally different from the existing drugs and therefore will have the potential to overcome existing resistances. As malaria is a disease of poor and developing countries, cost effective technologies have to be used to find the novel and potential entities. DNDi identified three potential gaps in the research and development of new drug development for malaria and other neglected diseases.

1. Discovery of novel targets and novel lead compounds. (Driven by public sector)

2. Clinical trials on validated drugs. (Has to be driven by pharmaceutical Industries)
3. Registration issues, lack of production, high prices (unaffordable by poor people)

It is very important to recognize and understand that parasitic drug discovery differs from chronic drug discovery process (preventable diseases such as diabetes, cancer, cardiovascular diseases, respiratory diseases etc are termed as chronic diseases [8]), not in terms of drug development process, but in terms of investment. Altruistic approaches and philanthropic institutions are needed to correct this fatal imbalance. WISDOM, which stands for "Wide *In silico* Docking on Malaria" is one such initiative that has been started as an altruistic approach to deal with malaria. The main goals and strategies employed in WISDOM project are described below.

Goals of the WISDOM project

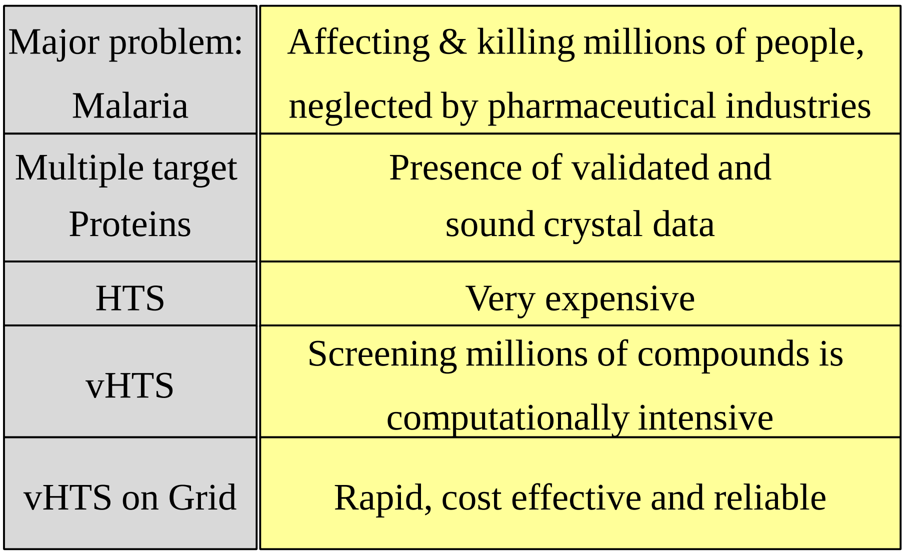
The main objective of the WISDOM project is to establish a collaborative framework between bio-informaticians, biochemists, pharmaceutical chemists, biologists, and Grid experts in order to produce and make selected lists of potential inhibitors against malaria and other neglected diseases. The main goals of WISDOM project are:

a. Biological goal: Identify inhibitors against malaria and other neglected diseases to be tested in the experimental laboratories
b. Grid goal: To develop a fault-tolerant WISDOM production environment that is capable of deploying molecular docking and molecular dynamics application or any other biomedical application efficiently on a Grid infrastructure.

This thesis mainly deals with the biological goals of the WISDOM project. The biological goals are dependent on the Grid goal, because, to achieve the biological goal a sustainable Grid infrastructure should be available. The Grid goal, which is the development of the WISDOM Grid production environment, is achieved in collaboration with our partners in the WISDOM collaboration.

Strategies employed in WISDOM project

Discovering hits with the potential to become usable drugs is a critical first step to ensure a sustainable global pipeline for discovery of innovative antimalarial products. While the establishment of public-private partnerships has helped to stimulate product R&D for some neglected diseases, increased emphasis needs to be placed on the high-risk early discovery phase. Hence, in the WISDOM project and in the current thesis, the focus is on discovery of new chemical leads; to achieve this, cost effective, reliable and robust *in silico* drug discovery methods are utilized. Figure 6 illustrates the rationale behind each strategy utilized in WISDOM project.



Major problem: Malaria	Affecting & killing millions of people, neglected by pharmaceutical industries
Multiple target Proteins	Presence of validated and sound crystal data
HTS	Very expensive
vHTS	Screening millions of compounds is computationally intensive
vHTS on Grid	Rapid, cost effective and reliable

Figure 6: Strategies employed in WISDOM project.
This Figure demonstrates the motivation, problems, and techniques employed in WISDOM project (on the left hand side). The reason why these techniques are used is described on the right hand side.

Drug discovery and *in silico* technologies

Hit identification is the first and foremost step in the drug discovery process [40]. Two different methods are widely used in the pharmaceutical industry for finding hits are high throughput screening and virtual screening [41]. In high throughput screening (HTS), the chemical compounds are synthesized, and physically screened against protein based or cell based assays. This process is commonly used in all major pharmaceutical industries. However, the cost in synthesis of each compound, in vitro testing and low hit rate are posing huge problems for pharmaceutical industries. Current efforts within the industry are directed to reduce the timeline and costs. Besides that, HTS campaigns to identify compounds causing a desired phenotype or entire pathways, many of these drugs are failing in clinical development either because of poor pharmacokinetic characteristics or to intolerable side effects, which may reflect insufficient specificity of the compounds [42]. At present, hundreds of thousands to millions of molecules have to be tested within a short period for finding novel hits, therefore, highly effective screening methods are necessary for today's researchers.

In view of the above problems in finding new drugs by HTS; cost effective, reliable *in silico* screening procedures are in practice. Especially *in silico* methods fit nicely when dealing with diseases such as malaria mainly due to their cost effective character. Hence, *in silico* methods such as virtual screening and molecular dynamics methods are used in the current thesis. A detailed description of the entire drug discovery process is given in Chapter 2.

Virtual Screening by molecular docking

Virtual screening provides a complementary or alternative solution to HTS in hit identification [43]. Such screening comprises innovative computational techniques designed to turn raw data into valuable chemical information and this chemical information into drugs. The definition of pharmacophores, pharmacophore searches, docking and scoring are currently well established in *in silico* drug design, giving new dimensions to this approach [44]. When structural information of the target protein is available, structure based methods are widely utilized. When physically compared to classical high throughput screening of chemical compounds, *in silico* screening is much faster and yields 10-100 fold higher hit rates at reduced cost [45]. Some of the more recent successful examples in rational drug design are the design of nonpeptide cyclic ureas for HIV protease, discovery of inhibitors for thymidylate synthase and inhibitors for acetylcholinesterase (AChE) [46, 47, 48].

Molecular dynamics methods

Due to the robust nature of docking algorithms, they in general ignore important parameters like protein flexibility and electrostatic solvation effects. This gap is filled with the more sophisticated molecular dynamics methods, which are based on force field calculations. Docking combined with molecular dynamics methods have been shown to be successful in several cases [49]. More detailed information on the drug discovery processes and the role of *in silico* methods are provided in detail in chapter 2.

Grid enabled molecular docking and molecular dynamics

The downside to vHTS is that screening millions of chemical compounds and rescoring the best hits by molecular dynamics is computationally intensive. The approach has a high computing and storage demand, therefore, it is termed as computational data challenge. Screening and further simulating each compound, depending on structural complexity, can take from one to a few minutes on a standard PC, which means screening a database with millions of chemical compounds can take years of computation time. Hence, modern concept of distributed computing termed as Grid computing is utilized. Computational Grid infrastructures are the best attempt to solving this problem thus far [50]. Computational Grids are a part of e-Science infrastructure that provides access to geographically distributed compute resources around the world. These resources range from personal computers to clusters of computers/super computer that belongs to several organizations. Generally, these compute resources are connected by using Internet protocols. Detailed description of Grid computing is given in chapter 3. The combination of these techniques (vHTS, molecular dynamics and Grid computing) can definitely decrease the financial cost implications of rational drug design strategies. Several docking applications have already been run on Grids and, proved to be successful. Some of the success stories in *in silico* drug design on computational Grids are the small pox research Grid [51], Anthrax research project [52] and Cancer project [53, 54]. The Grid technologies employed in the current thesis are described in detail in chapter 3.

Aims of the current thesis

This thesis is a part of the WISDOM project which aims at employing low cost *in silico* methods in combination with modern information technologies such as Grid computing for the identification of potential new cures for malaria. More precisely, this thesis mainly aims at predicting easily synthesizable small molecules against several targets implicated in malaria.

Besides that, the specific objectives of this thesis are:

a. To demonstrate how modern technologies such as Grid computing are utilized to accelerate the overall drug discovery process and deployment complex workflows on computational Grids.
b. To demonstrate how virtual screening by molecular docking is carried out on Grid to identify novel inhibitors against several targets of malaria.
c. To demonstrate how the combination of molecular docking and molecular dynamics simulations enabled hit identification.

1.2 Thesis outline

After giving the current state of the art on the neglected diseases and introduction to malaria biology in this chapter, the further chapters in this thesis are organized as follows:

Chapter 2 introduces the state of the art in molecular modeling techniques, with the special focus on the structure based drug discovery methods. It also gives an overview of the various algorithms and models that are used in *in silico* drug discovery with a particular spotlight on algorithms and scoring functions employed in this work

The role of molecular dynamics simulations in *in silico* drug discovery and descriptions of the general molecular dynamics simulations techniques are given in detail. The theory behind the molecular mechanics, molecular dynamics simulations and free energy calculations and the role of solvent are described in detail.

Chapter 3 introduces Grid computing and further describes the need of Grid computing in the life science area. Significance of computational Grids in the biomedical sciences research arena is described in detail with a special focus on Grids related to the drug discovery process. Finally, chapter 3 focuses on the role of computational Grids in the thesis. Further, the EGEE Grid infrastructure and the WISDOM production environment, which is designed with a special purpose to deploy the docking and molecular dynamics simulations, are described.

Chapter 4 focuses on the set up of molecular docking experiment in detail. This chapter explicitly describes the virtual screening effort against plasmepsin (part I of WISDOM project), with a special focus on the protein target involved, chemical compound database selection, validation, experimental setup, strategies in results analysis, docking results.

Chapter 5 focuses on the rescoring of the compounds selected from the molecular docking and in vitro results of the best 30 compounds selected. This chapter exclusively describes the impact of rescoring the docking conformations by MM-PBSA and MM-GBSA scoring functions. Finally, the modeling aspects of the final hits are described in detail. To confirm the identified hits as inhibitors against plasmepsin; inhibitory assays were performed by a laboratory in the WISDOM consortium, the methods used in this experiment and the results are described in detail.

Chapter 6 focuses on the docking experiment in which four different targets of malaria are screened against 4.3 million compounds from the ZINC database (Part II of the WISDOM project). The screening techniques employed were similar to the one described in chapter 4. The docking experiment outlined in this chapter follows a new multi-target approach.

Chapter7 summarizes the achievements and novelty of this thesis. This chapter also discusses the use and significance of current work in the area of academic drug discovery research and the role of collaborative research to deal with malaria. Finally, it provides conclusions and an outlook from the perspectives of the achievements in this work.

2 Chapter 2. State of the art on rational drug design

Computational methods are increasingly in practice in the drug discovery process and are very useful in hit and lead identification and further in lead optimization. This chapter introduces the general drug discovery process employed in biopharmaceutical companies with a special spotlight on rational drug discovery methods such as virtual screening by molecular docking and molecular dynamics methods.

This chapter is organized as follows: firstly, in section 2.1 the general drug discovery process is described with special focus on hit identification by high throughput screening and virtual screening. In section 2.2 virtual screening is discussed with focus on molecular docking. Advantages and disadvantages of various docking algorithms and scoring functions are described in detail. The state of the art on molecular dynamics methods with focus on minimization and free energy calculations is detailed in section 2.4. Finally, the use and significance of combining molecular docking and molecular dynamics in the identification of novel hits is described.

2.1 Drug discovery

Identifying or discovering novel drugs is defined as drug discovery (DD). DD whether driven by computational methods or experimental methods is a complex, challenging and multidisciplinary effort. Several phases of the drug design include discovery phase, optimization phase, clinical trial phase, registration, and approval by regulatory authorities (Figure 7). Besides its complexity, drug discovery is an extremely time consuming and expensive endeavor, it is estimated that the time and cost to bring a new drug to the market vary from 7-12 years and ~$800 million - $1billion respectively [55, 56, 57]. Figure 7 describes the different steps of drug discovery process and its associated costs. Though as shown in Figure 7, drug discovery is not linear workflow, it is a rather an iterative process. The aim of the process depicted in Figure 7 is to demonstrate the costs and time associated in identifying new chemical entities and further developing them into drug candidate molecules [49].

Discovery phase is the initial phase of the drug discovery process, which includes identification of disease, selection & validation of target and hit & lead identification. After target identification and validation, screening of chemical compounds is performed to identify the hits and leads. In the next steps, these hits and leads are further optimized in the process called lead optimization. The optimized leads enter into clinical trials phases. Finally, the drug has to be registered and approved by FDA or related organizations in other countries before entering the market [55].

Screening is the one of the first and foremost steps, careful and smart screening will lead to the identification of valuable hits, which later can be transformed into leads and drugs [40]. In pharmaceutical industries, generally two main screening techniques are employed: experimental screening also termed as high throughput screening (HTS) and virtual screening or *in silico* screening [41].

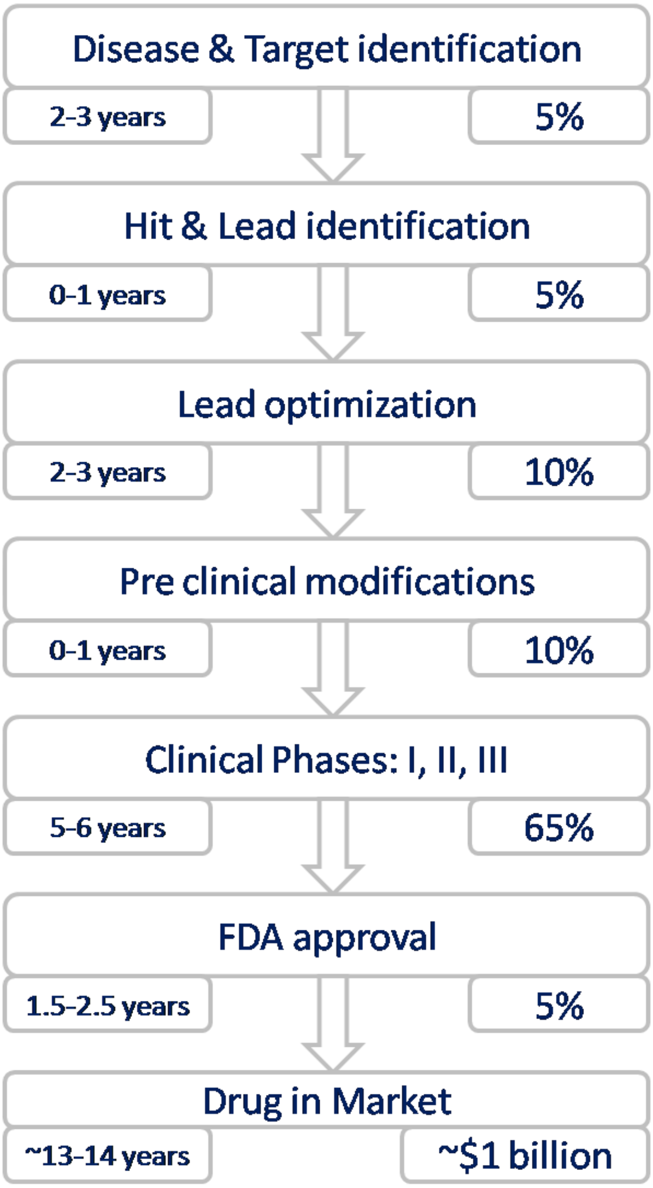


Figure 7: Classical drug discovery (DD) process employed in the pharmaceutical industries. The Figure illustrates several stages of DD process along with the approximate duration of time (on the left hand side) and percentage of total expenses involved in each stage (on the right hand side). Also demonstrates the total time and expense involved bringing a drug into market.

High throughput screening (HTS)

HTS is currently the central technique employed in larger pharmaceutical companies for finding the hits and leads. Screening of chemical compounds physically/experimentally against target protein is termed as HTS. Sophisticated, modern ultra fast robotic methods, which are capable of screening thousands of chemical compounds are currently available and are generally practiced in almost all the pharmaceutical industries [42].

In the initial steps of HTS, bioassays are to be setup, chemical compounds have to be synthesized (or can be purchased from chemical vendors), then screening and subsequent data analysis is performed. In the final steps, chemical compounds with high potency are identified and structure and mechanism of action are determined. However, determination of mechanism of action is still a question by HTS [42]. Though it is currently the main stream of screening chemical compounds in pharmaceutical industries and biotech companies, HTS is not without limitations. Some of the major constraints of HTS are:

- Cost in synthesis of each compound, in vitro testing, waste disposal, and low hit rate.
- False positives and unspecific binding of the tested compounds.
- Low solubility and non-specific reactions with the protein material, which results in surface adhesion or protein precipitation.
- From the knowledge point of view, HTS could not answer the question, why and how the detected hit acts upon the target.

Current efforts within the pharmaceutical industry are directed to reduce the time line and costs [58]. One alternative or complementary approach to HTS is, screening compounds by using rational drug discovery methods such as virtual high throughput screening [57]. Figure 8 illustrates the gain in hit rate using *in silico* screening over traditional HTS.

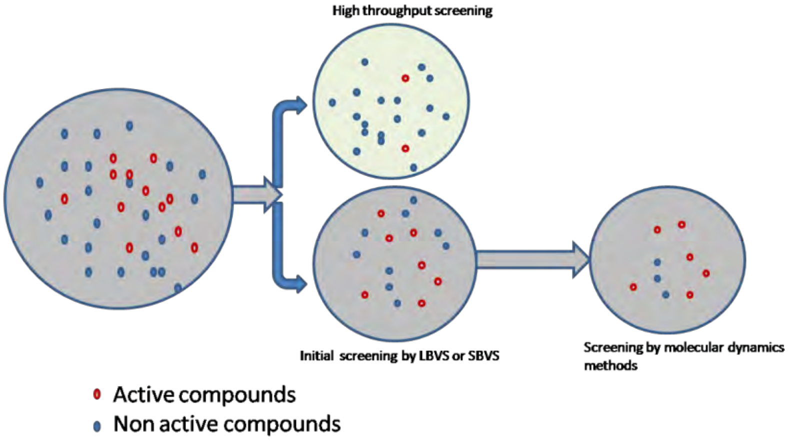

High throughput screening
Initial screening by LBVS or SBVS
Screening by molecular dynamics methods
○ Active compounds
● Non active compounds

Figure 8: Illustrates the increase in hit rate by using rational methods over random HTS. The Figure illustrates that using rational drug discovery methods will increase the hit rate when compared to random high throughput screening approach.

Why computer aided drug discovery

Besides the significant costs and time associated in bringing a new drug to the market, some of the major reasons for the pharmaceutical industries to look for alternative or complementary methods to experimental screening are [40]

a. Late stage attrition of chemical compounds in drug development and beyond [40]. Which in general is five of the 40,000 compounds tested in animals reach human testing and only one out of five reaching the clinical trials is finally approved [56].
b. Tremendous increase in chemical space and target proteins/receptors, this increases the demands put on the HTS and this in turn will call for new lead identification strategies (rational approaches) to curb costs and efficacy.
c. Advances in computing technologies on software and hardware enabled reliable computational methods

Computer aided drug discovery

According to Hugo Kubinyi [59], most of the drugs in the past were discovered by coincidence or trial and error method, or in other words, serendipity played an important role in finding new drugs. Current trend in drug discovery is shifted from discovery to design [59], which means, understanding the biochemistry of the disease, pathways, identifying disease causative proteins and then designing compounds that are capable of modulating the role of these proteins has become common practice in biopharmaceutical industries. Both experimental and computational methods play significant roles in the drug discovery and development and most of the times run complementing each other [41]. Rational drug discovery or computer aided drug discovery (CADD) is defined as a process by which drugs are designed/discovered by using computational methods. The main aim of the CADD is to bring the best chemical entities to experimental testing by reducing costs and late stage attrition. CADD involve [56]:

1. Computer based and information extraction methods to make more efficient drug discovery and development process
2. Build up chemical and biological information databases about ligands and targets/proteins to identify and optimize novel drugs
3. Devise *in silico* filters to calculate drug likeness or pharmacokinetic properties for the chemical compounds prior to screening to enable early detection the compounds

which are more likely to fail in clinical stages and further to enhance detection of promising entities.

There are various computational techniques, which are capable of affecting at various stages of the drug discovery process [44]. It is estimated that, computational methods could save up to 2-3 years of time and $300 million [57]. The two major disciplines of CADD, which can manipulate modern day drug discovery process and capable of accelerating drug discovery are, bioinformatics and cheminformatics. Figure 9 illustrates the impact of different rational methods in terms of time and cost on the drug discovery process. In general,

- Bioinformatics techniques hold a lot of prospective in target identification (generally proteins/enzymes), target validation, understanding the protein, evolution and pylogeny and protein modeling [43].
- Cheminformatics techniques hold lot of prospective in storage, management and maintenance of information related to chemical compounds and related properties, and importantly in the identification of novel bioactive compounds (hits and leads (NCE)) and further in lead optimization. Besides that, cheminformatics methods are extensively utilized in *in silico* ADME prediction and related issues that help in reduction of the late stage failure of compounds [44].

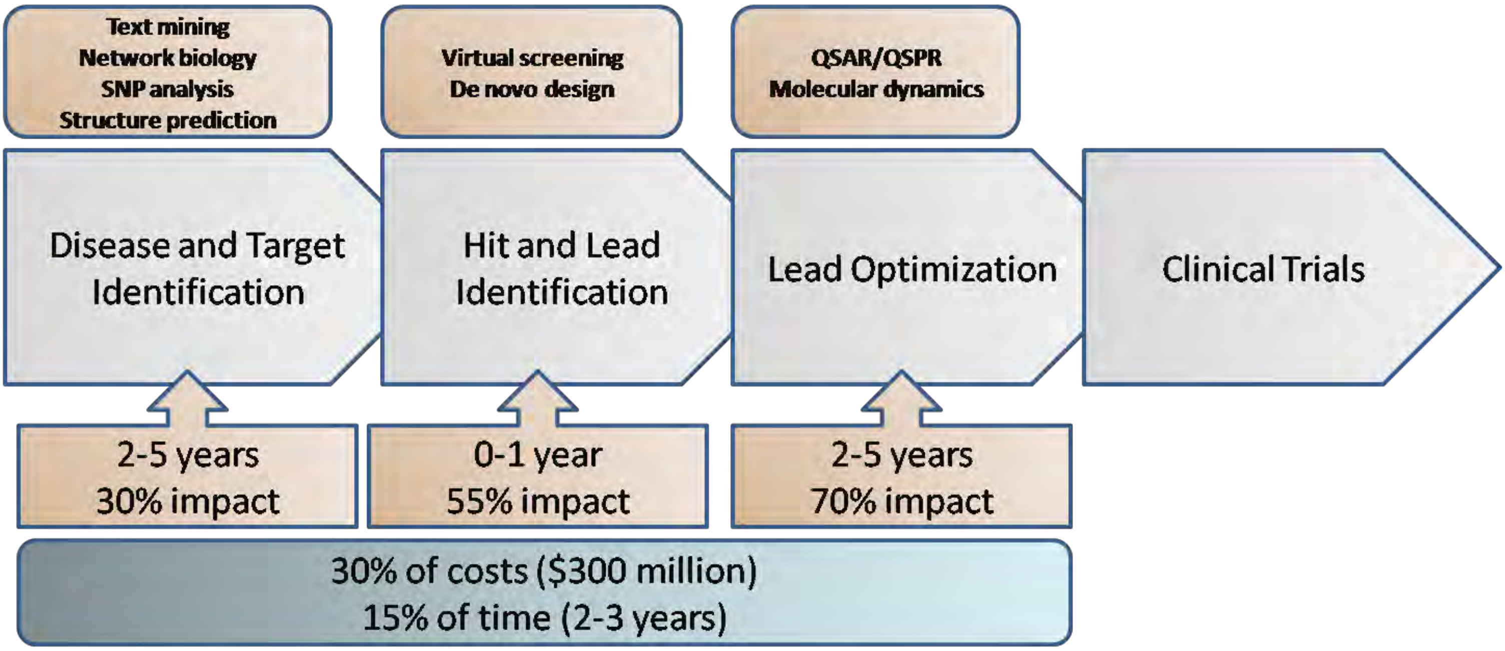


Figure 9: Illustrates the impact of rational approaches at various stages of the drug discovery process in terms of costs and time [60].
This Figure illustrates that a total of ~30% of the total costs and 15% of time can be saved by utilizing rational approaches.

In context to the current thesis, cheminformatics methods, especially techniques related to hit & lead identification and lead optimization are further discussed.

2.2 Virtual screening

In silico screening of chemical compound databases for the identification of novel chemotypes is termed as Virtual Screening (VS). VS is generally performed on commercial, public or private 2-dimensional or 3-dimensional chemical structure databases. Virtual screening is employed to reduce the number of compounds to be tested in experimental laboratories, thereby allows for focusing on more reliable entities for lead discovery and optimization [61, 62, 63, 64]. The costs associated to the virtual screening of chemical compounds are significantly lower when compared to screening of compounds in experimental laboratories. Virtual screening methods are mainly driven by the availability of the existing knowledge. Depending on already existing knowledge on the drug targets and potential drugs, these methods fall in mainly in these two categories (see Figure 10) [65, 66, 67]:

i. Structure based virtual screening or structure based drug discovery (SBVS or SBDD)
ii. Ligand based virtual screening (LBVS)

In the absence of receptor structural information and when one or more bioactive compounds available ligand based virtual screening are generally utilized. Different LBVS methods include:

a. Similarity search: Similarity searching is performed, when a single bioactive compound is available. The basic principle behind similarity searching is similar compounds have similar bioactivities.
b. Pharmacophore based virtual screening: When one or several bioactive compounds are available, pharmacophore based virtual screening is performed. The principle behind the pharmacophore is a set of chemical features and their arrangement in 3-Dimensional space is responsible for the bioactivity of the compound. By utilizing the these chemical features of already known bioactive compounds, a pharmacophore model is built, which later is used to screen against database of unknown compounds for finding chemical compounds with similar chemical features.

Similarity search methods, pharmacophore based methods [68] and ligand based virtual screening in general are reviewed in [69].

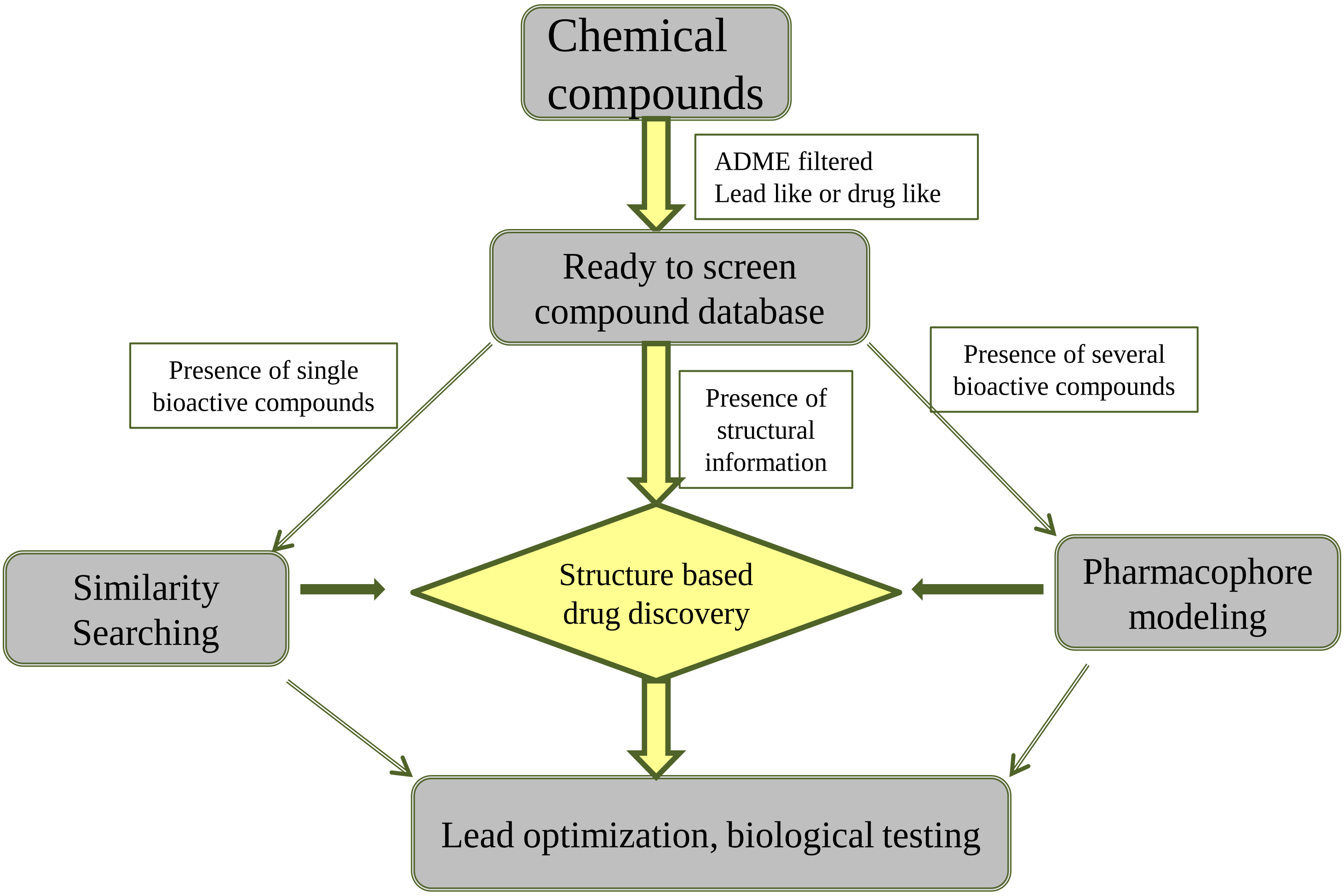

Chemical compounds
ADME filtered
Lead like or drug like
Ready to screen compound database
Presence of single bioactive compounds
Presence of structural information
Presence of several bioactive compounds
Similarity Searching
Structure based drug discovery
Pharmacophore modeling
Lead optimization, biological testing

Figure 10: Schematic representation of virtual screening methods [70].
The Figure illustrates the existence of various *in silico* screening methods, further it demonstrates the usage of these methods depending on the available data.

In the presence of structural information of the target protein, receptor based or structure based methods are widely used method to screen the compounds. Depending upon the availability of structural information, the screening can be performed by either using X-ray crystal models or NMR models or homology models. In context to the current thesis, SBVS methods are described in detail.

Structure based drug discovery

Structure based drug discovery methods (SBDD) are widely used in both pharmaceutical industry and academic institutes for finding novel chemotypes [58, 48]. SBDD uses knowledge of the target protein's structure to select candidate compounds with which it is likely to interact. Drug targets are usually most important molecules concerned in an explicit metabolic or cell signaling pathway that is known, or believed, to be related to particular disease state. Drug targets are most often proteins and enzymes in these pathways [71]. SBDD methods rely on the known 3D geometrical shape or structure of proteins for finding novel compounds. X-ray crystallography or nuclear magnetic resonance (NMR) techniques are typically employed to solve and obtain 3D structures of proteins/receptors. The capability of X-ray and NMR methods to resolve the structure of proteins to a resolution of a few Angstroms (about 500,000 times smaller than the diameter of a human hair) enabled researchers to precisely examine the interactions between atoms in protein targets and atoms in potential drug compounds that bind to the proteins. This ability to work at high resolution with both proteins and drug compounds makes SBDD one of the most powerful methods in drug design [71]. There are several examples for the successful application of SBDD methods, some of the recent successful examples in rational drug design are the design of nonpeptide cyclic ureas for HIV protease, the discovery of inhibitors for thymidylate synthase and inhibitors for acetylcholinesterase (AChE) [46, 47, 48].

Factors influencing the growth of SBDD

The major factors influencing the impact of SBDD methods are [72]:

- Advances in molecular biology, proteomics techniques: recombinant expression makes the isolation of large amounts of proteins much easier than before.
- Advances in X-ray crystallography and NMR techniques: Determination of the3-D crystal structures of proteins and receptors were made possible
- Advances in combinatorial chemistry and cheminformatics: Lead to tremendous increase in the chemical space and their availability in 2D/3D electronic databases.
- Online web services such as Brookhaven database [www.pdb.org]: Hosting of and providing structural information on thousands of disease related proteins/receptors enabling better understanding of protein-ligand interactions.
- Grid computing: Lead to perform data intensive scientific tasks easier than before. Further, it enabled sharing of terra bytes of scientific data between the research organizations.
- Availability of efficient and reliable molecular modeling, computational chemistry and result analysis tools.
- Finally, the availability of free resources such as ready to dock chemical compounds, web services, open source docking tools.

The general workflow

The prerequisite to set up a virtual screening experiment is knowledge on the target, against which the screening has to be performed, and on the chemical compound libraries. Most of the information related to the targets is available in the literature, whether it is digital or paper based. A typical virtual screening workflow involves the following steps and is illustrated in Figure 11

Step 1 Selection: Selection of the target, the chemical compound database, and the docking software.

Step 2 Preparation of the target: If the selected target is an X-ray crystal structure with a bound ligand, then it requires preparing the binding site of the protein by taking ~6–8 °A from the co-crystallized ligand, taking care the significant amino acids for the activity are included in the binding site. Information on the significant amino acids can be obtained either from the literature or from the Brookhaven protein database. However, both target and compound have to be prepared according to the needs of the software.

Step 3 Preparation of the compound library: After selecting the chemical compound database, one has to filter and remove undesired compounds. Lipinski's "rule of five" is one of the frequently used filters applied before the virtual screening campaign is started.

Step 3 Screening: Depending upon the number of compounds to be screened, one has to check for the availability of resources. If thousands of compounds are to be screened, distributed computing or Grid computing are utilized.

Step 4 Result analyses: Results are analyzed usually based on the docking score (free energy of the complex) and binding mode of the compound inside the binding site. Access to data analysis and visualization software is required at this point.

Step 5: Rescoring: Best scoring compounds are rescored by using sophisticated scoring functions.

Step 6 Selection: Visualization of interesting protein-ligand complexes, and final selection of compounds for experimental testing.

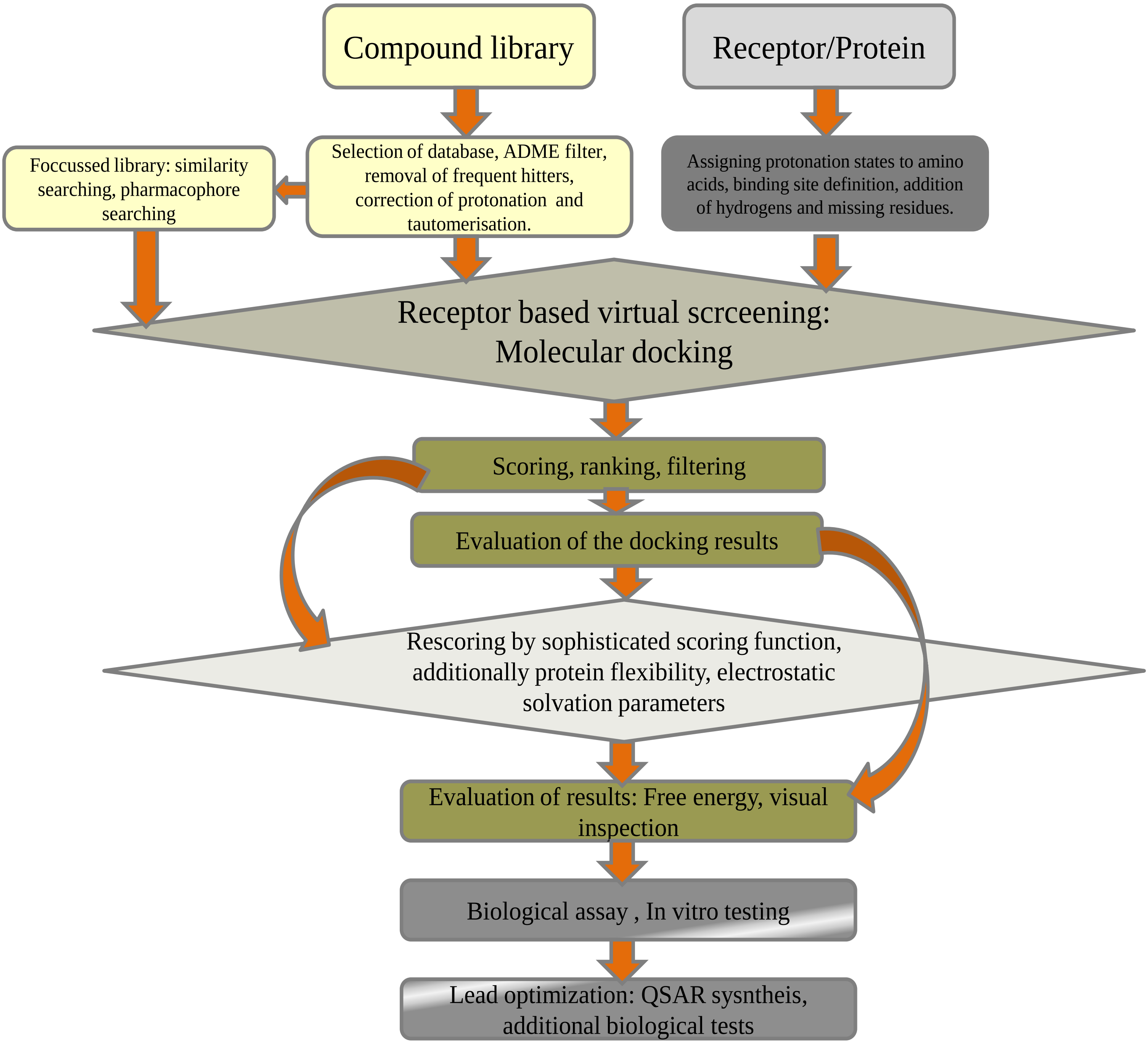

Compound library
Receptor/Protein
Foccussed library: similarity searching, pharmacophore searching
Selection of database, ADME filter, removal of frequent hitters, correction of protonation and tautomerisation.
Assigning protonation states to amino acids, binding site definition, addition of hydrogens and missing residues.
Receptor based virtual scrceening: Molecular docking
Scoring, ranking, filtering
Evaluation of the docking results
Rescoring by sophisticated scoring function, additionally protein flexibility, electrostatic solvation parameters
Evaluation of results: Free energy, visual inspection
Biological assay , In vitro testing
Lead optimization: QSAR sysntheis, additional biological tests

Figure 11: General receptor-based virtual screening procedure.
The Figure demonstrates the hierarchical virtual screening workflow starting from the chemical database preparation and receptor preparation to identification of novel compounds.

Goals of receptor based drug design

In the field of structure based drug discovery there are three major goals that theoretical biologists seek to achieve [73].

1. Accurately predict the conformation and orientation of both the protein and ligand in complex.
2. Rank order a database of chemical compounds against a particular target protein.
3. When only protein target knowledge is available, the goal is to discover or design novel chemical compounds, which are capable of inhibiting the target protein.

However, all the goals mentioned above are inter-linked to each other. Predicted binding energies or score of series of compounds cannot be achieved without predicting their binding orientations. Practical applications of these methods include finding novel compounds either by screening a database of compounds or de novo design of novel compounds by utilizing the features of protein active site. Additional applications include improving the binding of existing inhibitors (lead optimization) [73]. Receptor based methods such as molecular docking and molecular dynamics simulations are described below.

2.3 Molecular docking

The methodology of protein-ligand docking is inherited from earlier work in small molecule conformational sampling and macromolecular energy calculations [74]. Calculating the accurate protein-ligand interactions is the key principle behind structure based drug discovery [75]. Predicting ligand conformation within the active site of a protein/receptor is termed as molecular docking. In general, there are two key components of molecular docking [70]:

a. Accurate pose prediction or binding conformation of the ligand inside the binding site of the target protein.
b. Accurate binding free energy prediction, which later is used to rank order the docking poses.

The docking algorithm usually carries out the first part of the docking (predicting binding conformation) and the scoring function associated with the docking program carries out the second part that is binding free energy calculations.

Pose prediction: Docking algorithm usually perform pose predictions. Identifying molecular features, which are responsible for molecular recognition or pose prediction are very complex and often difficult to understand and even more so, when simulated on a computer [76].

The challenges and difficulties in the protein-ligand docking are mainly due to the involvement of many degrees of freedom

- Translational and rotational (in relation to each other) involves six degrees of freedom
- Conformational degrees of freedom (Protein and ligand)

The challenge of various docking algorithms lies in sampling these translational, rotational and conformational degrees of freedom accurately and further finding the ligand conformation which best matches the receptor conformation. Furthermore, the sampling has to be quick enough to allow evaluating number of compounds in a given a docking experiment [76].

Activity prediction: After the pose prediction by the docking algorithm, the immediate step in the docking process is activity prediction, which is also termed as scoring. Docking score is achieved by the scoring functions associated with the particular docking software. Scoring functions are designed to calculate the biological activity by estimating the interactions between the compound and protein target. During the early stages of the docking experiments, scoring was performed based the simple shape and electrostatic complementarities. However, currently, the docking conformers are often treated with sophisticated scoring methods that include the Van der Waals interactions, electrostatic interactions, solvation effects and entropic effects [77]. Detailed descriptions about different algorithms and scoring functions are given further this chapter.

2.3.1 Search methods and docking algorithms

Depending on the flexibility of protein and ligand, docking algorithms can be divided in 3 types:

- Rigid docking: Protein and ligand are considered to be rigid
- Semi-flexible docking: protein is fixed and ligand is flexible
- Flexible docking: Both protein and ligand are flexible

Based on the principle of conformation generation, the search methods are categorized into

- Stochastic
- Systematic
- Deterministic

Systematic search methods

Systematic methods attempt to investigate all the degrees of freedom (Both rotational and translational), but this may lead to combinatorial explosion. Hence, systematic methods utilize fragmentation/construction algorithms. Incremental construction algorithm [78] is the most commonly used systematic search method. Incremental construction of ligands in the active site of the receptor/protein can be accomplished in different ways; firstly, docking various ligand fragments in the active site of the protein and in the next steps, these fragments are linked covalently. This method is popular as de novo ligand strategy [76]. Alternatively, in another method, the docked ligands are initially fragmented into several fragments, the rigid fragments are considered as core fragments and the flexible parts of the ligands as side chains. In the next step, the rigid fragments are docked first and the side chains are later added to the rigid fragments incrementally in the binding site of the protein.

The basis for core fragment placement varies from one docking tool to another. It may depend upon the steric complementarity, as in the DOCK software or geometry based as in the FlexX software [78]. One of the major advances in the protein-ligand docking is the development of DOCK algorithm by Kuntz and co-workers, [79, 80, 81, 82] which is based on an “Anchor and build method”. In the DOCK program, the rigid fragment is placed based on the steric complementarity and the side chains are added to the rigid fragment one bond at a time in a systematic fashion exploring each bond’s conformational space [83]. To reduce the complexity of the problem, a pruning algorithm is used to remove the unfavorable conformations early on. FlexX [78, 84], SurflexX [85] are some of the other are widely applied docking programs, which are based on incremental construction algorithm. FlexX docking software is used extensively in the current thesis and is described in detail in chapter 4. There are several other algorithms which are in common to incremental construction algorithm but differ in the way rigid fragments are docked or the way flexible parts are added, for example, the Hammerhead algorithm [86, 87].

Stochastic methods

Stochastic search methods are also termed as random search methods. Stochastic search methods involve random changes to the position and as well as torsion angles for the ligand or pool of ligands to generate different conformations. The two most popular stochastic methods are genetic algorithm (GA) and Monte Carlo algorithm (MC) [88, 89, 90]. The Monte Carlo method is capable of generating ensembles of conformations statistically consistent at room temperature. While generating the pool of random conformations, with each iteration of the process, either the internal conformation of the ligand (by rotating around a bond) is changed or the entire ligand is subjected to the rotation or translation within the active site of the protein. An energy function evaluates the newly formed conformation and accepts the conformation only if the energy is lower than the one derived from the previous step or if, it is higher, is within the range defined by Boltzmann factor [87]. LigandFit [91], a popular docking program from Accelry's, is based on the Monte Carlo algorithm. GA starts with a population of random ligand conformations with random orientations and at random translations. In genetic algorithm (GA), each chromosome in a population encodes for one ligand conformation along with its orientation in its binding site of the protein. Then, in the next step, a scoring function evaluates the fitness of each individual in a population and less fit individuals are being killed (or not passed on to the next generation). Pairs of survived individuals are mated leading to children with new chromosomes derived from the parents by mutations and recombination. (Chromosome in this context refers to position, orientation, and conformation of the ligand). GA differs from the Monte Carlo methods by performing a number of runs and selecting the structures with highest scores. GOLD [92], AutoDock [93] and DARWIN [94] are the some of the few docking programs, which rely on genetic algorithms.

Deterministic methods

Deterministic methods are also termed as simulation methods include molecular mechanics and molecular dynamics methods. Unlike the systematic and stochastic methods, deterministic method address the issue of both ligand and protein flexibility. By the definition itself, in the deterministic search, the initial state determines the change that can be made to generate the next state, which generally has to be energetically preferred as compared with the initial state. One of the major limitation of simulation methods is the trapping up of the ligand conformations in local minima (local minima corresponds to conformation with higher energy than the low energy stable conformation) on the energy surface rather than stable low energy global minima, this is because deterministic methods cannot cross the high-energy barriers within feasible simulation time [76]. Longer simulation time can be one solution, but cannot be adapted to virtual screening of thousands of ligands. One strategy to overcome the local minima problem is by starting the molecular dynamics simulation by using different ligand positions. In distinction to molecular dynamics simulations, molecular mechanics methods often reach local minima only and are not used as standalone methods, but rather complement other methods, for example in the Monte Carlo method within the DOCK program [87]. Some simulation methods that could be useful for VLS have been developed to overcome the energy barriers more rapidly, for example using simulated annealing molecular dynamics implemented in SDOCKER program [95].

In context to virtual screening, molecular mechanics and molecular dynamics are mostly recommended at the final steps of the hierarchical virtual screening process on the preselected smaller library of the compounds derived from docking experiments (Figure 11). Workflows starting with an initial screening of compounds by robust docking algorithms, followed by sophisticated simulation methods are widely recognized, and proposed by several authors [96, 97].

2.3.2 Scoring functions

One of the two important components of molecular docking is scoring. While docking aims at reproducing binding conformation close to the X-ray crystal structure, the aim of scoring is to quantifying the free energy associated with protein and ligand in the formation of the protein-ligand interactions. Most of the docking software are associated with scoring functions, which enable computing free energy associated with protein-ligand interactions (docking score). The docking score is used to rank the chemical compounds in virtual screening campaign. Wide ranges of scoring functions are available to calculate the binding between the protein and virtual ligand. These methods range from estimating the binding by simple shape and electrostatic complementarities to the estimation of free energy of protein and ligand complex in aqueous solutions. Only few of them are capable of addressing the thermodynamic process (described further in this chapter) involved in the binding process. However, methods based on thermodynamic parameters require extensive simulating time, and consequently significant CPU time, therefore, these methods are restricted to smaller set of compounds, making it impractical to use them in large-scale virtual screening experiments. Currently three main types of existing scoring functions are applied: Force field-based, empirical scoring functions and knowledge based scoring functions [98]. A short description on each scoring function is given below. Detailed information about the scoring functions is reviewed in [42, 76].

Force field based scoring functions relies on the molecular mechanics methods. Force field-based methods calculate both the protein-ligand interaction energy and ligand internal energy and later sums both the energies. Different force field functions are based on different force field parameter sets. For example, AutoDock relies on the Amber force field and G-Score relies on the Tripos force field [98].

The Van der Waals and electrostatic energy terms describe both the internal energy of the ligand and the interactions between the protein and ligand. The van der Waals energy term is described by the Leonard Jones potential and often can be varied depending upon the desired hardness of the potential. Electrostatic terms are described by the Coulombic formula with a distance dependent dielectric constant for charge separation. Advantages of force field based scoring functions include accounting of solvent and disadvantages include over estimation of binding affinity [98] and arbitrarily choosing of non bonded cutoff terms [76].

Knowledge based scoring functions: Atom pair interaction potentials also known as potential of mean force (PMF). Atom pair interaction potentials are usually derived from structural information stored in databases (ChemBridge structural database and protein data bank) of protein-ligand complexes. It relies on the assumption that repeated occurrence of close intermolecular interactions between certain types of functional groups or atom types are energetically favorable than the randomly occurring interactions, thus contribute complementarily to the binding affinity. The robust nature of this scoring function makes it usable in virtual screening experiments. Knowledge based scoring function rely on existing intermolecular interaction databases, one major limitation of this method is the limited availability of such structural information in the intermolecular interaction databases. Dscore [99] and PMF scoring functions rely on knowledge based scoring functions [100].

Empirical Scoring functions: The **s**core in empirical scoring function is derived from individual energy contributions of each component involved in intermolecular interactions. Empirical scoring functions are easier to apply and subjected to less computational error. For example, the Kuntz ID, in his early work emphasized on the molecular shape, because shape complementarity is certainly essential for a ligand to be placed in the binding site and can be easily and accurately computed. However, in his later work he added chemical information, molecular mechanical energies, and empirical hydrophobicities to make the scoring function more accurate [79, 81]. Boehm HJ developed an empirical scoring function that takes into account hydrogen bonding, ionic interactions, lipophilic contact surface and number of rotatable bonds [101, 42]. Due to its robust nature, empirical scoring functions are widely used in virtual screening experiments along with knowledge base scoring functions. One of the major limitations of empirical scoring function is that it works very well with rigid ligands, but the results are not satisfying with flexible ligands. This is because most of the empirical scoring functions ignore the internal energy of the ligand. FlexX (docking tool) and Ludi (de novo design tool) [101] rely on empirical scoring function.

Given the significance of structure based drug discovery, especially protein-ligand docking, currently there are several docking software available [98]. These tools are developed either by commercial bioinformatics companies or by research institutes. Table 3 illustrates the some of the major docking tools utilized in pharmaceutical industries and academic research institutes.

Docking Tool	**Algorithm/Method**	**Scoring function**	**Flexibility**
FlexX www.biosolveit.de	Incremental construction / SS	Boehm empirical scoring function	Protein: No Ligand: Yes
FlexX-Pharm www.biosolveit.de	Incremental construction / SS	Boehm empirical scoring function	Protein: No Ligand: Yes
AutoDock http://AutoDock.scripps.edu	Simulated Annealing and Genetic algorithm / StS	Force field based empirical scoring	Protein: No Ligand: Yes
Dock http://dock.compbio.ucsf.edu	Incremental construction/ SS	Force field based scoring	Protein: No Ligand: Yes
ICM http://www.molsoft.com/	Simulated Annealing / StS	Force field based scoring	Protein: No Ligand: Yes
GOLD http://www.ccdc.cam.ac.uk/	Genetic algorithm / StS	Empirical knowledge based scoring	Protein: Partial Ligand: Yes
Surflex-Dock http://www.optive.com/	Incremental construction / SS	Empirical Hammerhead scoring	Protein: No Ligand: Yes
Glide http://www.schrodinger.com	Simulated annealing & Incremental search / SS & StS	Empirical knowledge based scoring	Protein: Yes Ligand: Yes

Table 3: Illustrates widely used docking tools.
This Table demonstrates the existence of several docking tools. The algorithms and scoring functions used by these docking software are given exclusively. In the last column on the right hand side, information on the ability of the software to consider protein and ligand flexibility is provided.
SS = Systematic search, deterministic search = DS, Stochastic search =StS

Limitations and Challenges

Although various algorithms and scoring functions exist to solve the docking problem, SBDD is not without limitations. Some of the major limitations of structure based drug discovery are [49, 66, 76, 98]: (listed according to the priority)

1. **Protein flexibility:** Even when the structure of the target molecule is known, the ability to design a molecule that binds to inhibit or activate the target remains a major challenge. Although the fundamental goals of virtual screening methods are to identify those molecules with the proper complement of shape, hydrogen bonding, and electrostatic and hydrophobic interactions for the target receptor, the complexity of the problem is far greater in reality. For example, the ligand and the receptor may exist in a different set of conformations when in free solution, which is different from the conformation when the ligand is bound to a protein [61].
2. **Role of solvent and scoring functions:** Proteins and ligands are usually surrounded by solvent; typically water molecules. The entropy of the unassociated ligand and receptor is generally higher than that of the complexes, and favorable interactions with water are lost on binding [98, 61]. These energetic costs of the association must be offset by the gain of favorable intermolecular protein–ligand interactions. The magnitude of the energetic costs and gains is typically much larger than their difference, and, therefore, potency is extremely difficult to predict. Even though several methods have been developed to predict the strength of molecular association events accurately by accounting entropic and solvation effects [102, 103], these methods are costly in terms of computational time. Thus, make them inappropriate to use in screening of large compound databases.
3. **Increase in the Biological and Chemical data:** Advances in genomics research area led to the rise in X-ray resolved structures and made them accessible through online web services, such as protein databank (www.pdb.org). The possible chemical space is currently estimated to be in the range of 10^{60} chemical compounds [104]. Although this astronomical number is not synthetically possible; advances in combinatorial chemistry lead the way to synthesize millions of chemical compounds in timely manner [72]. Both the rise in biological data and chemical data is normally considered as positive aspects for virtual screening, but the real challenge is, it requires huge computing resources to screen these compounds against several of targets proteins.

4. **Data analysis tools:** Data analyses methods are of great importance especially in view of the very large data generated, because of large-scale approaches such as virtual screening. However, unavailability of customized docking databases that enable storing and sorting millions of records is a significant limitation for large-scale experiments.

The challenge in developing practical virtual screening methods is to develop an algorithm that is fast enough to rapidly evaluate potentially millions of compounds while maintaining sufficient accuracy to successfully identify a subset of compounds that is significantly enriched in hits [63, 64, 76]. Several groups have attempted to utilize various scoring functions to address the errors arising from single scoring functions; this process is termed as consensus scoring [105, 106, 107, 108, 109]. Though, better hit rate and success was achieved by using consensus scoring [107, 110, 111], the issues: protein flexibility, role of the solvent in mediating protein-ligand interactions and accurate prediction of binding free energy still remain unsolved by the docking methods currently available.

2.4 Molecular dynamics

It is widely accepted opinion [49] that docking results need to be post-processed with more accurate modeling tools before biological tests are undertaken. Molecular dynamics (MD) simulations has great potential at this stage:

A. Firstly, it enables a flexible treatment of the ligand/target complexes at room temperature for a given simulation time, and therefore is able to refine ligand orientations by finding more stable complexes. (Achieved by **Minimization techniques).**

B. Secondly, it partially solves conformation and orientation search deficiencies that might arise from docking. (Achieved by **Molecular dynamics simulations).**

C. Thirdly, it allows the re-ranking of molecules based on accurate scoring functions. (Achieved by **Free energy calculations).**

A. Minimization

Flexible treatment of protein and ligand can be performed by a method known as minimization. Minimization is a process by which the molecular structure is brought to the minimum energy conformation. Minimization involves iteratively adjusting atomic coordinates until the forces acting on all the atoms in the system become zero or close to zero.

Minimization generally takes the molecule to the local minimum, nearest to the starting conformation.

The energy of the molecule is a useful property to study the behavior of the molecules. Provided sufficient parameters such as how long the bond is, how strong the bond is, molecular mechanics models can be used to approximate the energy of the molecules. However, one should be aware of the fact that minimization does not find lowest energy conformation rather they find unstrained molecular conformation [112]. Minimization of the protein, ligand or protein-ligand complex can be achieved by molecular mechanics and quantum mechanics methods. Accurate energy calculations can be achieved by using molecular orbital theory based quantum mechanical methods. However, they in general are slow and require significant computing resources, thus make them unsuitable for large-scale virtual screening experiments.

Energy surface

Changes in the energy of the molecule can be considered as changes on a multidimensional surface called energy surface. The stationary points on the energy surface are particularly interesting, where the first derivative of the energy is zero with respect to internal or Cartesian co-ordinates. At stationary point, the forces acting on all atoms are zero. Minimum points are one type of stationary points that corresponds to relatively stable structures.

In molecular modeling, the minimum points on the energy surface are interesting; because minimum energy arrangements of the atoms correspond to the stable state of the molecule and any movement away from a minimum gives a conformation with higher energy and thus makes it relatively less stable conformation than the previous conformation.

There may be very large number of minima (minimum points) on the energy surface. The minimum with very lowest energy is termed as global minima. The highest point on the pathway between the two minima is of special interest and is known as the saddle point. Both minima and saddle points are stationary points on the energy surface, where the first derivative of the energy function is zero with respect to all the co-ordinates [112].

Minimization algorithms are used to identify the geometries of the molecules that correspond to the minimum points on potential energy surface. Based on the potential energy calculation on the energy surface, minimization algorithms are classified into two groups: derivative method and non-derivative method. The most commonly used methods in molecular modeling of drugs are derivative methods: steepest descent and conjugate gradient methods.

Minimization by Molecular mechanics

Empirical methods such as molecular mechanics ignore the electronic motions and calculate the energy only based on the nuclear positions only. This makes molecular mechanics methods suitable for calculation of system with large number of atoms (for example systems such as Protein-Ligand complexes). Though molecular mechanics is an empirical method, in some cases, it provides as accurate solutions as similar to quantum mechanical calculations in relatively less time. However, molecular mechanics fails to provide the properties that depend on the electronic distribution of the molecule [44].

Molecular mechanics calculations or force field calculations are based upon a simple model of the interactions within a system with contributions from processes such as bond stretching, opening, and closing of bond angles and rotation around the single bonds. Force field is defined as a mathematical function that describes the potential energy of the system. The main components of force field are covalently bonded terms and non-covalently bonded terms. Covalently bonded terms are bond lengths, bond angles, and dihedral angles. Non-covalent bonded terms are the van der Waals terms and electrostatic terms [112]. The complete function typically resembles the following equation:

$$E_{MM} = \sum E_{bond} + E_{angle} + E_{torsion} + E_{vdw} + E_{electrostatic}$$

This summation, when given in explicit form, represents a force field evaluating the potential energy as a function of the geometry. The most commonly used force fields are AMBER [113], Charmm [114] and Gromacs force fields [115]. Amber force field is utilized exclusively in the current project and is described in detail in chapter 5.

Application of molecular mechanics methods:

Energy minimization is widely used in molecular modeling and is an integral part of techniques such as conformational search procedures. Energy minimization may be used prior to molecular docking, molecular dynamics simulations, and the Monte Carlo simulation in order to relieve any unfavorable interactions in initial conformation of the system.

B. Molecular dynamics

Molecular dynamics simulations (MD) are widely applied and are an accepted computational technique for studying biological macromolecules. Molecular dynamics simulations are extensively used in understanding protein folding, refining docked conformations, calculating accurate free energies and entropies. Besides the use of MD simulations in rationalizing the experimentally derived properties at molecular level, one of the major applications of MD simulations is the refinement of the experimentally derived X-ray crystal and NMR structures. The relationship between molecular dynamics techniques and experimental techniques is longstanding, with theoretical methods (MD) help in understanding and analyzing the experimental data. In turn, experimental methods help in validation and improvement of computational methods [116, 117].

Theory

Biological molecules are studied at microscopic and atomic level by using molecular dynamics simulations. Molecular dynamics method describes and calculates the time dependent behavior of biological systems. The simulation begins by giving each atom in the molecule some kinetic energy. This makes the molecule to move around, and it is possible to calculate, how the molecule moves by solving the Newtonian equations of motion. In molecular dynamics, integrating Newton's laws of motion generate successive configurations of the system. The result is a trajectory that specifies how the positions and velocities of the particles in the system vary with time [112]. The trajectory is achieved by solving the differential equation embodied in the Newton's second law of motion (F=ma)

$$d^2x_i/dt^2 = Fx_i/m_i$$

In context to the current thesis, molecular dynamics simulations are extensively utilized to solve three problems, firstly, to remove the conformational and orientation deficiencies that are raised from docking methods, secondly, to address the issue of protein flexibility and finally, to calculate accurate binding free energies between protein and ligand.

C. Free energy calculations in protein-ligand interactions: Thermodynamic cycle

The stable structures of a small molecule correspond to minimum points on the multidimensional energy surface, with alternative conformations populated according to their free energies. The free energy of a particular conformation is equal to the solvated free energy at the minimum with the small correction of configurational entropy about the minimum point (stable structure) [118]. The Anfinsens renaturation experiments [119], showed that this basic principle of statistical physics also applies to protein [120]. Thus, although there are specific proteins that become trapped in the local minima, because of the barriers to folding or covalent bonding, most of the proteins fold to the conformation of minimum free energy [120]. The same argument should apply to reversibly binding ligands, therefore, it seems reasonable to assume that small molecule ligands adapt the binding mode of the lowest free energy within the binding site of the protein. This very reasonable assumption is the ultimate basis for the use of energy in the molecular docking experiments.

Ligand binding to the protein is often achieved by the non-bonded interaction such as hydrogen bonding, hydrophobic or lipophilic or aromatic contacts. Typically, these parameters are enough when predicting biological activity.

If reliable/accurate binding free energy of a protein-ligand complex is to be achieved, then complex thermodynamics aspects have to be taken into account. Binding free energy or Gibbs free energy is a result of both entropic and enthalpic contributions. Usually, protein and ligand (when exist as single entities) are surrounded by solvent, typically water molecules. When these proteins and ligands are in solvent, they frequently make interactions with the surrounding water molecules. These interactions between the protein/ligand with the solvent are rearranged, when the ligand binds to the protein (desolvation). The energetic parameters determining these interactions have to be considered, but most importantly, the breaking or formation of any new interactions will be related to the changes of ordering parameters of the entire system. The ordering parameters are termed as entropic terms. Hence, when equilibrium conditions are considered (Complex = protein + Ligand, in solvent), in addition to the hydrogen bonding, hydrophobic, steric, internal strain, factors such as desolvation and entropic contributions (both rotational and traslational) are important and have to be into account. More detailed thermodynamics parameters determining protein-ligand binding is given in [42]

Free energy calculations in virtual screening

In context to virtual screening by molecular docking, techniques that are capable of correctly predicting the ligand conformation in the active site of the protein, and accurately ranking the final ligand conformer are necessary. In general, free energy calculations in virtual screening by docking have two major challenges. Firstly, it should be able to effectively discriminate the best conformation from the pool of conformations of the same system (because docking produces many docking poses for the same ligand), and secondly, it should correctly predict the relative stability of different complexes (because in virtual screening, many ligands are screened). While insight in the relative stability of different complexes is sufficient in the initial screening experiments, but estimation of absolute free energy is essential in the later stages of docking, particularly in lead refinement, where only a few selected complexes are considered. Different molecular dynamics based calculations can be carried out on the docked complexes to estimate the correct binding free energies as well as accurate ranking [121, 122, 123, 124].

Thermodynamic integration and free energy perturbation methods are the most stringent methods available for calculating the binding free energies. However, the requirement of exhaustive sampling to reach convergence and the computational expensive nature of these methods makes them unsuitable for large-scale virtual screening purposes [125, 126].

Molecular dynamics based calculations such as Molecular Mechanics- Poisson Bolzmann surface area calculations (MM-PBSA) [127] and Linear Interaction Energy (LIE) calculations [128, 129, 130, 131] provide relatively good binding free energies in reasonable time and at moderate cost. This enables these techniques suitable for virtual screening experiments. The MM-PBSA method is exclusively utilized in the current thesis, and is described in detail in chapter 5.

2.5 Combination of docking and molecular dynamics methods

The need to combine docking and molecular dynamics methods stems from the underlying weakness and strengths of both the methods. While strengths of molecular docking include, robust screening of chemical compounds and fast conformational sampling, weaknesses include ignorance of solvent parameters and protein flexibility. In contrast to docking, molecular dynamics methods strengths include, protein flexibility and inclusion of solvent parameters and the weakness include longer simulations times and significant CPU resources [49]. Hence, by combining these two techniques, by using molecular docking to initially screen a database of compounds robustly and then treat the few selected complexes by molecular dynamics methods to include protein flexibility and calculation of accurate free energies is a valid approach and has a great potential to find new lead compounds.

Success stories by combining docking and molecular dynamics methods

The two step protocol, molecular docking for initial screening and in the next step applying molecular dynamics simulations on the selected protein-ligand complexes appears to be a practical approach to address the structure based virtual screening problem. Some of the successful examples [49], where molecular dynamics simulations were utilized to optimize docking conformations, are: rationalizing the inhibitor specificity of CDK2 [132], discrimination of stable and unstable human acetylcholinesterase ligand conformations [133], discovery of novel inhibitors against ALR2 [134], and the generation of robust QSAR model from the final structures of steroid complexes [135]. Other studies include the origins of the enantioselectivity of an antibody catalyzed Diels-Alder reaction [136], interaction modes of nimesulide and prostaglandin-endoperoxide synthase-2, [137] and optimization of the manually docked structures of several glucocorticoids within a model of the glucocorticoid receptor [135].

2.6 Summary

In silico drug discovery activities gained remarkable significance in the past few years and are becoming inseparable from the experimental drug discovery activities. The role of rational methods either in accelerating or in enhancing drug development is immense, especially in terms of cost and time. This chapter summarizes state of the art on rational drug discovery methods currently employed in pharmaceutical industries and academic research area. Both, ligand-based and structure-based drug discovery methods are extensively used in identifying hits. Structure-based drug discovery methods work well, when structural information of receptor is available, particularly, when the receptor structure is available with a co-crystallized ligand/substrate.

Docking is the method of first choice for rapid *in silico* screening of large ligand databases for drug research, since it is based on a rational physical model and very fast. However, there is very often a compromise between speed and accuracy of the results (in terms of the actual binding mode as well as the calculated affinity values) concerning the best scoring docking solutions. Most probably, among the many number of possible conformations that a ligand may adopt within the binding site of the receptor, a quite good one will be generated indeed, but not necessarily ranked among the first few predictions due to the approximate nature of the scoring function.

Even though qualitative or quantitative consensus scoring addresses the problem of false positives and false negatives to a little extent, addressing solvation parameters and protein flexibility are still major issues by the docking methods. Thus, it seems reasonable to subject the docking predictions to force field calculations, which provide more detailed energy and charge models, and thus, allows us to focus only on the most reasonable predictions. Once the stage of force field calculations is set, further structure optimization becomes feasible for the ligand, as well as for the receptor. Finally, the free energy between ligand and receptor is computed.

This chapter highlights the need of combining docking with molecular dynamics and potential advantages of this combination. The current thesis proposes a hierarchical workflow, which starts with robust screening of chemical compounds by molecular docking and post process the selected complexes by means of complex molecular dynamics based simulations that enable protein flexibility and accurate free energy calculations. However, screening large databases by molecular docking and further optimizing the protein-ligand complexes by molecular dynamics is computationally expensive, thus requires significant computational resources. To overcome this problem, computational Grids are utilized and are described more in detail in chapter 3.

3 Chapter 3. Deployment of molecular docking and molecular dynamics on EGEE Grid infrastructure

This chapter introduces the modern concept of distributed computing termed as Grid computing. The main aim of this chapter is to introduce Grid computing, and to discuss its impact and significance in modern day drug discovery process. Later in this chapter, in context of the WISDOM project, the deployment of molecular docking and molecular dynamics applications on EGEE Grid infrastructure is discussed[1].

3.1 Introduction

Several definitions of Grid

Ian Foster and Carl Kesselman, pioneers of the Grid, proposed a definition in 1998: "*A computational Grid is a hardware and software infrastructure that provides dependable, consistent, pervasive, and inexpensive access to high-end computational capabilities*" [50].

In 2000, with Steve Tuecke, they added to the original statement; Grid computing is concerned "*with coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organizations*" [138]. Then in 2002, Ian Foster modified his definition again, arguing: the Grid is "*a system that coordinates resources that are not subject to centralized control, uses standard, open, general purpose protocols and interfaces, delivers non-trivial qualities of service*" [139]. The computing Grid name comes from the well-known analogy with an electrical power Grid [140]. The computing power would be delivered just like electricity from an outlet, without knowing where the power came from or its complexity and reliability. An obvious similarity between computational and electrical Grids is that both aggregate heterogeneous power sources (thermal, hydro, or nuclear power and workstations, clusters or supercomputers) [140]

3.1.1 Concept of e-Science

e-Science (Enhanced-Science) term is described as the large-scale, distributed and collaborative science that is enabled by the advances in the Internet technology. Currently, scientific research is more and more carried out by communities of researchers that span disciplines, laboratories, organizations and national boundaries. The main goals of e-Science

[1] This chapter is based on Kasam V, Salzemann et al. Large-scale Deployment of Molecular Docking Application on Computational Grid infrastructures for Combating Malaria. ccGrid, pp. 691-700, Seventh IEEE International Symposium on Cluster Computing and the Grid (CCGrid '07), 2007

are to support these interdisciplinary research collaborations – including those cross-institutional boundaries. Information and computing technologies play a vital role to carry out such interdisciplinary, large-scale collaborative science. Though World Wide Web provides us the information on the web pages wherever on the Internet, mere accessing of web content in most cases is not enough for performing effective scientific research. A more sophisticated and secured infrastructure is needed to enable e-Science, which allows the scientists to access not only the content on the web but also the underlying machinery such as computing resources and/or directly the databases that are storing the information.

Computational Grids are a part of e-Science infrastructure that provides computer resources to carryout various types of research. Grids are supposed to handle several types of applications and data in a secured and reliable way, in this sense they are ideal workbenches for e-Science.

3.1.2 Computational Grid

Computational Grids are built by the gathering and sharing of geographically distributed computing resources, typically but not necessarily clusters of computers [50]. The main motivation behind the development of a computational Grid is to use the freely available and unused computing resources, either belonging to an organization or personal computers for scientific research. The general concept of Grid is building a unique fault-tolerant system whose resources are accessible by several users (organized in virtual organizations, VO (the concept of VO is explained in detail further in this chapter)) all at once, in a transparent way. Computational Grids gather as many resources as possible, so that at one point, the Grid itself is enough to satisfy the user’s needs in terms of computing power, storage space, and further guarantee the security and confidentiality of data through secured authentication, authorization and replication. In the ideal Grid infrastructure, the physical locations of resources do not matter anymore as the applications and data have logical references to these redundant distributed locations. This differs from the Internet where the user has to choose to which machine he wants to connect and which information he wants to retrieve out of the tremendous amount of data available [141]. In this context, single points of failure do not exist anymore, and no part of the Grid is truly critical enough to threaten the availability and reliability of resources. Of course, these features can be achieved through several layers of software and services termed as middleware.

3.1.2.1 Grid architecture

Basic Grid architecture consists of four layers: Application layer, Middleware layer, Resource layer and Network layer. The Grid architecture demonstrated in Figure 12 is just a general model.

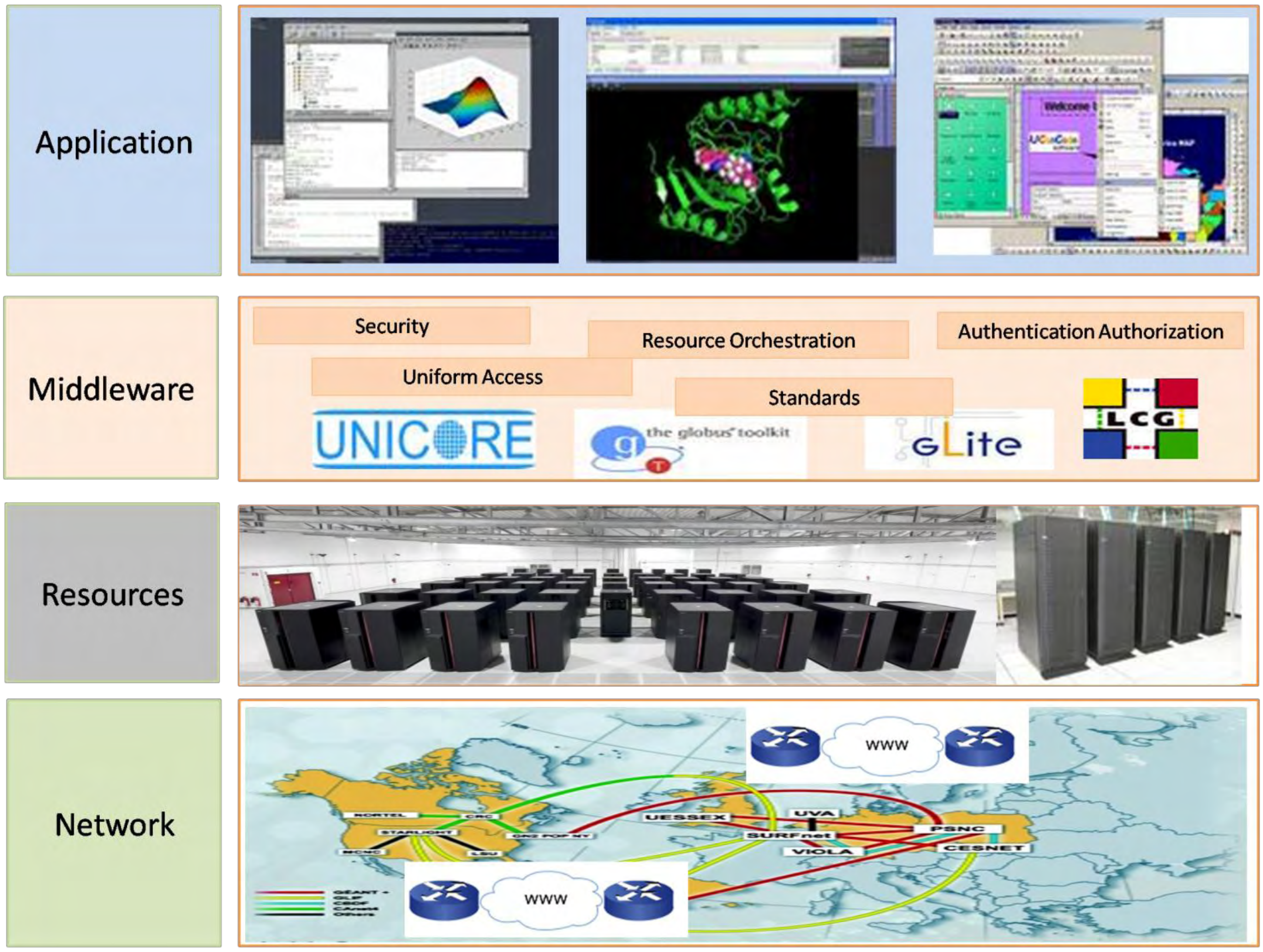


Figure 12: General Grid architecture [142]
The Figure demonstrates that there are there are four layers in Grid computing in general, Top layer constitutes Application layer, followed by middleware later and finally the resource layer and network layer connecting resources.

The application layer of the Grid describes different types of the applications that can be deployed on the Grids, the applications range from science, engineering, simulations, particle physics etc, and portals. These applications are usually developed using Grid-enabled programming environment and interfaces, and the services provided by the user-level middleware. The resource layer consists of the main computing resources (CPU), these resources are geographically distributed across various organizations and mainly controlled by the local resource managers. Network layer is the one that connects these geographically distributed resources (using Internet protocols). The layer between the resource and applications (the user) is termed as middleware. Grid middleware usually can be divided into two layers: user-level middleware and the core middle ware. . The core middleware layer offers services to abstract the complexity and the heterogeneity of the resource level. The user-level middleware layer provides higher-level abstractions and services. Grid middleware is usually infrastructure specific. It takes care of the distribution of the jobs, security issues, and status checking of the jobs on the Grid. In summary, Grid middleware is the key technology of the Grid computing. In context to the current thesis, EGEE infrastructure is utilized, which rely on gLite middleware. More details about the gLite middleware, its components, protocols, and services are provided later in this chapter.

3.1.2.2 Grid security and Virtual organizations

Security is an important aspect that concerns Grid. As resources in the Grid are provided and accessed by geographically distributed organizations and individuals. The deployment of applications such as drug discovery application where data security and integrity are most important, concerns Grid Computing. For example in a drug discovery application, the molecule data within chemical databases and experimentation results are often sensitive, and need to be protected. Effective protection of intellectual properties and sensitive information requires, for instance, authentication of users from different institutions, mechanisms for management of user accounts and privileges and support for resource owners to implement and enforce access control policies [143].

The Grid users are typically organized in virtual organizations. According to Ian Foster [138], a virtual organisation (VO) is defined as "an infrastructure that enables flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions and resources”. A VO is a Grid-wide identification and authorization unit representing a community of users sharing Grid resources. The main aim of members in a VO is sharing data, software and computational capabilities securely between the users of the organization. For instance, the biomedical virtual organization, which provide computational resources to perform biomedical research at large-scale). The resources in the virtual organization can be accessed only by the authorized users, and prevents the unauthorized users from accessing the Grid resources and computational capabilities.

To be able to join a virtual organization, and access the Grid, users must own a personal certificate, issued by a recognized certification authority. There is a consortium, EUGridPMA [144], which gathers all the recognized authorities. One should ask his certification authority for a certificate, and then register into a virtual organization. Once registered, the owner of a certificate, the user can access to the Grid through the user interface, then the user has access to a set of commands and APIs to interact with the Grid [145, 146, 143].

Though virtual organizations provide an overall security in restricting unregistered users from accessing the Grid resources; secured access to resources even within all the members of VO is needed.

In the Grid community, X.509 security specifications are usually implemented to handle security issues like authenticating users and servers by promising secure communication over the networks. Grid certificates based on the X.509 (Public/Private key) standards are used for the users and services (i.e. hosts) due to the mutual authentication process imposed by the Grid Security Infrastructure (GSI) [147]. These certificates are issued, and signed by trusted certificate authorities (CA). Authentication is typically done only once per connection between a client and a server.

3.1.3 Classification of Grids

Based on the hardware resource, Grids can be categorized into mainly two types: Desktop Grid and Cluster Grid. The sharing of personal computers that are connected by the Internet forms desktop Grids (inspired from peer-to-peer). As various PC users voluntarily donate the computing resources in the Desktop Grid, they are also termed as volunteer Grids. Some famous examples of volunteer Grid projects are Africa@home [148] and Seti@home [149]. On the other hand, Cluster Grids are formed by sharing of dedicated computing resources that are distributed across several organizations in the world. The EGEE grid infrastructure [150] is a well-known example that falls in the cluster Grid category.

A. Desktop Grids or Volunteer Grids

Low CPU usage in the desktop PCs' around the world has been reported in the last decade [151, 152, 153, 154]. Many of these cheap PCs' are connected to the Internet. As a result, several initiatives were planned to take benefit of this available computing power. A desktop Grid [155] is defined here as the sharing of idle desktop workstations or PCs, cycles to solve scientific problems. When the computing resources are freely available from the public through the Internet, the desktop grid is called a volunteer grid [156, 157]. SETI@home initiative [149] is the well-known example in volunteer computing, which utilizes BOINC platform [158] ((Berkeley Open Infrastructure for Network Computing). Millions of participants processed a database of large pulsar signals in a search for extraterrestrial intelligence. A central server to PCs distributes radio data signals through the Internet, where a screensaver program to assess the presence of a non-random signal that runs while the computer would otherwise be idle analyzes them. The notion of virtual organizations [50] and virtual enterprises [159] emerged from this success. The understanding that 10,000 desktop PCs with an average performance of 500 megaflops and appropriate software are equivalent to a 5 teraflops supercomputer developed a computational economy for sharing and aggregating resources to solve problems [154]. Consequently, there are many different desktop grid initiatives in life science [154].

Desktop Grids are highly used by different projects, but they are mainly used for computationally intensive projects. In contrast to desktop Grids, cluster Grids offers a larger variety of services.

B. **Cluster Grids**

A cluster is a set of computing units physically gathered in the same place and coordinated to improve computer capacity and storage. Clusters can have different sizes and can be composed of heterogeneous hardware. Two of the advantages of this approach are the scalability and the cost: a cluster can grow simply by adding new cheap PCs to it [143].

A cluster grid [160] is defined here as the sharing of geographically distributed clusters to solve problems. They allow selection and aggregation of distributed resources, such as instruments across multiple organizations. This enables exploration of large problems with huge data sets, but also small but daily needs. High-level services for complex applications can also be built on cluster grid [161]. They can be the infrastructure for data and knowledge grids, supporting collaborations and expertise. A cluster grid may interconnect hundreds of clusters supporting each day thousands of jobs. Examples of such Grids are EGEE [150], EELA [162], EUChinaGrid [163]. The European project, EGEE is the largest production cluster grid in the world and is adapted for computing power and data intensive applications.

Computational Grids can be classified in many ways. The more classical categorization of Grids is based on the application development or services provided by Grid. Based on this, Grids are classified into:

A. Knowledge Grids

Knowledge Grid is defined as setting that connects data and information in a transparent way and provides knowledge to the end users. Knowledge can be defined as the sum of all types of data and information within the scope of interest and is composed of relevant databases, information sources, document/knowledge bases, metadata, and a knowledge map. These

Grids are concerned with the way that knowledge is acquired, used, retrieved, published, and maintained to assist users in achieving their particular goals and objectives. In context to pharmaceutical research, knowledge Grid should support virtual laboratories (e.g. myGrid [164], Virtual lab [165, 161]).

B. Data Grids

Establishing secure access to distributed datasets and their management is the main aim of these Grids. To provide a scalable storage and access to the data sets, they may be replicated, catalogued, and even different datasets stored in different locations to create an illusion of mass storage [139]. The processing of datasets is carried out using computational Grid services and such a combination is commonly called data Grids. Sample applications that need such services are management, sharing, and processing of large datasets in high-energy physics experiments and accessing distributed chemical databases for drug design.

C. Computing Grids

Computing Grids are also termed as production Grids. They provide the resources to perform computationally intensive tasks in a transparent way by using automated job submission and distributed facility. Examples of computing Grids are EGEE [150], NASA IPG [166], the World Wide Grid [167], and the NSF TeraGrid [168]. The range of applications includes high-energy physics, particle physics, chemical engineering, and biomedical applications such as, homology modeling, molecular docking and molecular dynamics simulations.

3.1.4 Service oriented architecture and web services

Accessibility of the applications is an important issue in Grid computing. A need has emerged for communities to have standards and a standards-based architecture that would facilitate better interoperability among various grid middleware systems and Grid-enabled applications [169]. The Open Grid Services Architecture (OGSA) [170, 171] based on Web services Resource Framework [172] aims at unifying web and grid service frameworks. A Web service is a software system designed to allow inter-computer interaction over a network. Web services allow grid developers to take advantage of standard message formats and communications mechanisms for communicating between heterogeneous components and architectures. Recent developments in the Grid computing include wrapping the applications in Web service, this enable the user to conveniently access the applications on Grid from anywhere, if connected to the Internet [142]. The initial idea behind Web services is to enable the World Wide Web to become more and more the support for real applications and a mean for communication between them. The Web services specifications recommended by the

World Wide Web Consortium (W3C) propose a set of standards and protocols allowing interaction between distant machines over a network. These interactions are made possible through the use of standardized interfaces and protocols, which describes basically what are the available operations in a service, what are the messages exchanged (requests and responses), and where the service is physically located on the network and through which support. This interface is written in WSDL (Web Service Description Language) [173]. The WSDL specification itself leverages XML standards such as Simple Object Access Protocol (SOAP) for message exchange and XML Schemas for complex data structures.

The main advantages of Web services are [174]:

- They offer great interoperability (mainly because of standardised specifications).
- They enable the communication of processes and transfers of data independently of the programming language of the underlying applications. Therefore, by extension, virtually almost any piece of software can be exposed as a Web service.
- They can be considered as firewall-friendly, because they are based on standard Internet protocols.

The main weaknesses of Web services are:

- They are not adapted for transferring huge quantities of data.
- The performance can be worse with respect to other Remote Procedure Call (RPC) based communication methods due to the overhead of sending XML messages

Beside the above mentioned advantages, one prominent use of Web services in context to Grid is “virtualizing access to scientific applications”, which in other words mean simplifying the Grid usage by introducing the Web service between the user and Grid. This hides the complexity of the Grid (middleware, resources such as computers or super computers) and further provides an abstract interface to a given scientific application to be deployed on the Grid. Given the significance of Web services usage in Grid computing, the paradigm is currently shifting from infrastructure development to the development of Grid services, protocols and interoperability. This paradigm shift ensures that services are no longer restricted to particular Grid middleware and therefore, support multi-system, multi-language open architectures. Web services technologies have become standard for accessing applications on the Grid. Web services are usually used to build service-oriented architectures (SOA), as they are language and platform independent, and further enable access via multitude of interfaces. More details about the interoperability and service-oriented architectures are given in [175]. Some of the well-known Web service based life science applications on Grid are NPS@ [176, 177], which provide bioinformatics services such as protein sequence analysis, The DockingServer [178], which provides commercial virtual screening services, CHARMMing (CHARMM INterface and Graphics) [179], which provides molecular dynamics simulation services. More details about the web service based life applications can be found in our recent review [142].

3.1.4.1 Workflow technology

Workflow technology is another feature adopted by Grid community [142]. Diverse types of available resources such as databases, software applications, computing resources, services (Web service or Grid service) are integrated in a generic mechanism termed as "workflow technology". These workflow systems make possible the information exchange within different fields of life sciences, such as molecular biology, clinical chemistry, and computational life sciences. The entire idea behind workflow technology is to enable the researchers to develop their own protocols for scientific analysis by incorporating and accessing diverse distributed tools and data resources running on different hardware platforms, including web services and Grid computing [142]. Several applications of workflow technologies include drug discovery, genomics, large-scale gene expression analysis, proteomics, and system biology. The main advantage of using workflow technology is, it allows the life science researchers to perform the integration of different algorithms and tools without requiring any programming skills [180].

There are many workflow systems currently available; these include both commercial and open source systems. Scitegic Pipeline Pilot from Accelry's [181] and the InforSense platform [182] are the main commercial workflow system providers in the area of life sciences. Both of them support wide of operations related to cheminformatics, bioinformatics, and computational chemistry and further provides links to external databases and applications. Other workflow systems include KNIME [183] and TAVERNA [184]. While KNIME is freely available for non-profit organizations and for-profit for commercial organizations, TAVERNA, which is a part of ^{My}Grid project [164] provides complete open source services. TAVERNA [184] is closely integrated with Grid systems i.e., Condor [185] and its extensible architecture help easy integration of third party software tools.

So far, this chapter has introduced the concept of e-Science and Grid computing. In context to the Grid computing, the general Grid architecture, organization of the resources in the Grid i.e., virtual organizations, and Grid security were discussed. Relatively modern concepts in Grid computing: Web services and workflow systems, and their significance were also discussed in detail. In context to the current thesis, the following sections briefly discuss the importance of Grid computing in life science research with a special focus on Grids in virtual screening of chemical compounds.

3.2 Computational Grids in life sciences

Pharmaceutical industries have traditionally encouraged and adopted new technologies to enhance and accelerate the drug discovery process. Example of such adoption is, inclusion of rational drug discovery methods. The impact of rational methods, especially, structure based drug discovery (SBDD) methods, such as molecular docking and molecular dynamics on the drug discovery process is already demonstrated in chapter 2. SBDD can influence the process of new drug finding by reducing costs and time, but SBDD as a standalone technique cannot fetch new drugs into market. Drug discovery is synergy of several disciplines of biology, chemistry, molecular modelling, computational chemistry and computer science. In most pharmaceutical R&D environments, there is large amount of data and information generated by these various fields of science in search of drugs. For a drug discovery process to be successful, these data has to stored, properly managed, shared and analyzed. The main requirement at this stage is an IT infrastructure, which can manage tools that in turn can respond quickly to changing needs and, more importantly, enable rather than hamper the ability to innovate. Grid infrastructures have a lot of potential at this stage; they have the ability to manage, store, share and process huge volumes of data.

3.2.1 Biomedical applications on computational Grids

Biomedical applications that are utilizing distributed computing and Grid computing belong to relatively new area of research that was unveiled in the past 10 years, and are currently fully evolving. Table 4 gives an overview on current and recent biomedical projects using computational Grids. Example projects for each step in the drug discovery process (from discovery phase to Clinical trials) where Grid computing played a significant role are described in [142]. Even though computational Grids are capable of executing many types of life science applications, top area where computational Grids can in fact deliver value for research and development of new drugs in the pharmaceutical industry and academic research is, drug discovery. Computational Grids affect various stages of discovery process. These applications range from target identification to lead identification and validation, and clinical data management and sharing. In context to the current thesis, only lead identification with a special focus on Grid-enabled virtual screening is discussed.

Project Title	Website	Keywords
@neurIST	http://www.aneurist.org	Integrated biomedical informatics; individualized patient risk assessment
BIG GRID	http://www.nikhef.nl/Grid/BIG	the Dutch e-Science Grid
BioSapiens	http://www.biosapiens.info/	Integrated Genome Annotation
BRIDGE	http://www.bridge-Grid.eu/	Drug design scenario: virtual screening by several molecular docking tools
caBIG	https://cabig.nci.nih.gov/	The cancer Biomedical Informatics Grid (caBIG) connects individuals and institutions to enable the sharing of data and tools for worldwide cancer research.
CancerGrid (UK)	http://www.cancerGrid.org	Anti-cancer drug design
Cardioworkbench	http://www.cardioworkbench.eu/	Drug Design for Cardiovascular Diseases: Integration of *in silico* and in Vitro Analyses
D2OL	http://www.d2ol.com/	The Drug Design and Optimization Lab (D2OL)™ works to discover drug candidates against Anthrax, Smallpox, Ebola, SARS and other potentially devastating infectious diseases.
DataMiningGrid	http://www.dataminingGrid.org/	Data mining applications on standards compliant Grid service infrastructures; Grid-assisted re-engineering of gene regulatory networks and analysis of proteins using computational simulations
DEISA	http://www.deisa.eu/	Bio-molecular simulations, molecular docking
EMBRACE	http://www.embraceGrid.info	Integration of major databases and software tools in bioinformatics
EUMed-Grid	http://www.eumedGrid.org/	Empowering eScience across the Mediterranean, several bioinformatic tools
EuroGrid	http://www.euroGrid.org/	Bio Grid: Biomolecular simulations, structural analysis
GEMSS	http://www.it.neclab.eu/gemss/	Grid enabled medical simulation services
GridLab	http://www.Gridlab.org	Data management and visualization
OpenMolGRID	http://www.openmolGrid.org/	Speed up drug-design, ADME filtering, QSAR
SIMDAT	http://www.simdat.org	Pharma activity with data integration; distributed workflow tasks
ViroLab	http://www.virolab.org/	Individualized HIV treatment optimization; molecular dynamics
World Community Grid	http://www.worldcommunity Grid.org/	public computing Grid, runs FIGHTAIDS@HOME and Discovering Dengue Drugs – Together.

Table 4: List of recent and current biomedical applications utilizing computational Grids. These displayed Grid projects are in relevance to Biomedicine, especially drug discovery applications.

Lead Identification

The significance of *in silico* compound screening is already discussed in chapter 2. Large database of chemical compounds are routinely screened in pharmaceutical industry for finding novel chemotypes. Screening millions of chemical compounds in the computer brings along a high storage complexity, which means a computational data challenge on its own. Screening each compound, depending on structural complexity, can take from a few minutes to hours on a standard PC, which means screening all compounds in a large virtual compound library, can take years of computation time on a single machine. This problem can be addressed by distributing the workloads on a large computational Grid of thousands of computers, thus, reducing the time to screen the large virtual compound libraries to days. Figure 13 illustrates the need of using Grid computing while performing large-scale virtual screening experiment. Some recent successful examples of virtual screening on Grids are Screensaver Lifesaver project [54], FIGHTAIDS@HOME [148], and Discovering Dengue Drugs –together [186].

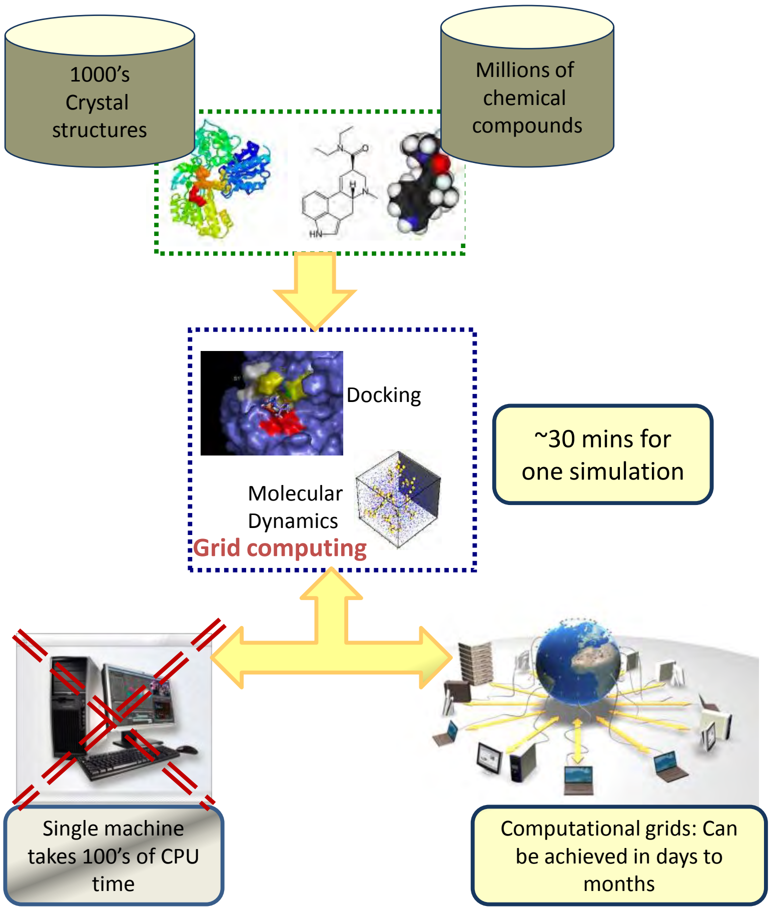

1000's Crystal structures
Millions of chemical compounds
Docking
Molecular Dynamics
Grid computing
~30 mins for one simulation
Single machine takes 100's of CPU time
Computational grids: Can be achieved in days to months

Figure 13 : Grid enabled virtual screening.
The Figure demonstrates that there are currently 1000 protein crystal structures and millions of chemical compounds available in databases. This enables virtual screening (docking and molecular dynamics), but screening all these compounds on machine is theoretically not feasible. Computational Grids provide an opportunity to deal with such CPU intensive applications.

Screensaver Lifesaver project [54] is one the first and most successful drug discovery application, where they used large-scale virtual screening approach to find novel inhibitors against several targets implicated mainly in cancer, and in anthrax and smallpox diseases. This project was inspired by SETI@HOME [149] (Search for Extraterrestrial Intelligence, described in Section 3.1.3). They used computing power donated by several volunteers all around the world (volunteer Grid computing), and screened 3.5 billion chemical compounds using cheminformatics tools THINK and LigandFit. Up to 10% of the predicted compounds were experimentally active in one specific case [187].

FIGHTAIDS@HOME [148], which used volunteer Grids and Discovering Dengue Drugs – together [186], which used cluster Grids are the some of the other Grid projects, which used structure-based drug discovery approaches to find inhibitors against AIDS and Dengue fever respectively.

Enterprise Grid

The usage of Computational Grids is not restricted to academic institutions or governmental organizations, they are successfully used in pharmaceutical industries; such a Grid is termed as enterprise Grid. This type of Grid is achieved by mutualisation of computing resources in an organization. Many pharmaceutical companies, such as Bristol-Myers Squibb and Novartis, are using idle time of thousands of desktop computers. They acquire teraflops of cheap computing power for their drug discovery research through enterprise Grid technology.

Example of such Grid is, computational Grid at Novartis Pharmaceuticals [188] that aims at finding novel anti-cancer agents [189]. United devices technology [53] is utilized to link 2,700 company PCs (Pentium 4 processor platform) to form a high performance-computing infrastructure. This type of Grid is accomplished by using the untapped processors when the workers in the company left their computers unused and/or when they are in meetings or left to home after end of the working day. Enterprise Grid computing technology significantly reduces the necessity to buy new computers. Novartis considerably enhanced their computing power by using enterprise Grid technology without adding any new computers to the existing ones, and performed a high-throughput docking experiment on 400,000 chemical compounds against human CK2 by using docking software, DOCK [190]. One very potent inhibitor that was ever reported before was discovered after post-processing the docking results. Computational Grid at Novartis described here is a typical example of Grid-enabled virtual screening application in pharmaceutical industry [189].

In summary, computational Grids are increasingly used in both academic and industrial settings to accelerate the drug discovery process that too at a reduced cost.

3.3 WISDOM – Wide *In silico* Docking on Malaria

In context to the current thesis, WISDOM project is described with a special focus on how molecular docking and molecular dynamics applications were deployed on EGEE Grid infrastructure. The following sections from now on will describe the EGEE Grid infrastructure, WISDOM production environment, the deployment of large-scale docking against four different target proteins implicated in malaria, and finally the deployment of molecular dynamics simulations against multiple target proteins implicated in malaria. The biological and molecular modeling aspects of WISDOM project are discussed in chapter 4, 5 and chapter 6.

3.3.1 EGEE

EGEE is a production Grid project that aims at building a Grid infrastructure for e-Science [150]. It was initiated by the needs of CERN [191] to process data coming from the largest machine in the world, the Large Hadron Collider (LHC) [192]. The project also developed its own middleware, gLite [193] that offers services to build a Grid. EGEE is a large Grid infrastructure built up from dedicated resources around the world, institutes, computing centers, laboratories etc. The resources range from simple desktop computers to clusters of computers. Currently EGEE is now the biggest Grid infrastructure in the world with more than 41000 CPUs and more than 10 Petabytes of storage. There are many services to keep the access transparent for the user (job submission and monitoring management, data management, information system etc.). The Grid is available to scientists 24 hours-a-day, 7 days a week; its use is thus flexible and the experiments can be easily reproduced [145, 194, 143]. The huge number of available resources on EGEE allows too many users to work together and to manage medium to large deployments; what would be impossible otherwise [145, 194, 143].

Services

Several protocols and services have been developed for gLite, a lightweight middleware for Grid Computing. The security mechanisms used by gLite are based on Grid Security Infrastructure (GSI). All users are identified by certificates, and Grid map files created in conjunction with the information registered on the VOMS servers [195] (Virtual Organization Membership Service), provide the authorization part, gLite is a middleware specifically developed for (Scientific) Linux (until 2007). Thus it is currently rather time consuming to deploy it under a different operating system but porting efforts to other operating systems are under way. The Grid system interoperates with the underlying batch queues. Main services and elements of the EGEE middleware are listed below:

Resource Broker: The resource broker is the main service for research selection and job submission. It handles jobs from their submissions by users to the retrieval of the results. It is also a scheduler, doing match making to send the job to the right resources and monitoring the job as it executes on the worker node. It allows for running applications on remote computing resources.

Computing Element: These services handle job execution and provide information on job characteristics and status. They actually host the batch server, all the worker nodes behind being the batch clients.

Storage Element: It is a service that allows virtualizing many types of storage; single or array disks, tapes servers, etc. Secure and reliable file transfers are mainly performed with GridFTP.

Single Catalogue or LFC (LCG File Catalogue): This central service registers files, replicas, and logical names. It has a unique catalogue to manage replicas and files. In the Grid, a Grid Unique Identifier (GUID) identifies each file, which is a unique identifier that links to a unique logical file name. One GUID can have several replicas, stored all over the Grid. Each replica is identified by a unique physical name. The catalogue supports VOMS and GSI for authentication and authorization.

Information System: The information system keeps track of both user related as well as Grid application specific metadata to discover resources and information about the resources.

Grid security and accounting: Authentication, authorization and auditing (AAA) is supported by gLite using the Grid Security Infrastructure (GSI (Public/Private keys)).

User Interface: This is the entry point of the Grid for the users. It is a set of command-line tools, GUIs, APIs to allow access to the main services of the Grid: Workload Management System and Data Management. User interface does the following operations:

a. Get information on the available resources
b. Submit a job and/or cancel a job
c. Get the status of the submitted jobs
d. Retrieve job results
e. Get information about the job (its submission and its execution)

f. Store files on the storage elements and replicate them, copy or delete files from the Grid.

Grid Infrastructures utilized in WISDOM project

Along with the EGEE Grid, the molecular docking and molecular dynamics deployments were achieved on several Grid infrastructures: AuverGrid [196], EELA [162], EUChinaGrid [163] and EUMedGrid [197]. All these infrastructures are actually using the same middleware, gLite. EGEE is the main infrastructure offering the largest resources; they are all interconnected with EGEE, in the sense that all of these Grids share some of their resources with EGEE. In the case of AuverGrid, it is even more evident as all the resources available through the AuverGrid Virtual Organization (VO) are also shared with several EGEE VOs.

3.3.2 WISDOM production environment for molecular docking and Molecular Dynamics

A large-scale deployment requires the development of an environment for job submission and output data collection. A number of issues need to be addressed to achieve significant acceleration from the Grid deployment. Grid performances are affected by the amount of data moved around at job submission. Therefore, the files providing the 3D structure of targets and compounds should preferably be stored on Grid storage elements in preparation for the large-scale deployment. The rate at which jobs are submitted to the Grid resource brokers must be carefully monitored in order to avoid their overload. The job submission scheme must take into account, this present limitation of the EGEE brokering system [143, 145].

Grid submission process introduces significant delays for instance at the level of resource brokering. The jobs submitted to the Grid computing nodes must be sufficiently long in order to reduce the impact of this middleware overhead.

Specific issues related to WISDOM deployment

A molecular docking and molecular dynamics job requires input files (target protein and chemical compound file), molecular docking/molecular dynamics software (Amber), and executable script to perform the desired experiment. A number of issues need to be addressed to achieve significant acceleration from the Grid deployment. Previous experience within our group with LCG middleware indicated potential bottlenecks [143]:

- **Reduction of input output operations:** The amount of data moved around at job submission affects the Grid performances. Thus, input-output operations play important role in Grid, hence, all the input required for either docking or molecular dynamics stored once on the storage element of the Grid and are retrieved when ever required.

- **Over load on Resource broker:** As the resource broker is the key element in distributing the jobs, care should be taken to avoid over loading. Over loading of resource broker may lead failure of all jobs passing through a particular resource broker.
- **Long jobs:** Grid submission process introduces significant delays for instance at the level of resource brokering. The jobs submitted to the Grid computing nodes must be sufficiently long in order to reduce the impact of this middleware overhead.
- Use of licensed software requires designing a strategy to distribute licenses on the Grid.

Description of the WISDOM production environment

WISDOM environment has been used two times in previous large-scale experiments, WISDOM-I [145] (against the plasmepsin family of proteins) in summer 2005 and a second deployment against avian flu in the spring 2006 [146]. Nicolas Jacq and Jean Salzemann at IN2P3-CNRS, France, developed the WISDOM production environment. Since its first deployment, the WISDOM environment keeps evolving in order to make it more users friendly, and easier to use by non-Grid experts. Jean Salzemann and I together performed the molecular docking deployment, which is described in the following sections

The main objective was also to improve the fault-tolerance of the system, in implementing, for instance, a persistent environment, that can be stopped and restarted at any time without risk of losing significant information. This proved to be also very useful as it enables the whole maintenance of the scripts and code, and improve the interactivity with the user. The user can also manage jobs finely, for instance force the cancellation and resubmission of a scheduled job. Along with this, we tried to minimize the cost of the environment in terms of disk space and CPU consumption for the user interface. Most of the job files are now generated dynamically: this allows the user as well to modify on the fly the configuration of the resource brokers and the jobs requirements. Through this, the user is sure that the next submissions will consider these modifications. Figure 14 demonstrates the overall architecture of the environment.

The user is interacting with the system through the two main scripts (widom_submit and WISDOM_status) deployed on the User Interface. These scripts will automatically take care of job files generation, submission, status follow-up and eventually resubmission. The jobs are submitted directly to the Grid Workload Management System (WMS), and are executed on the Grid computing elements and worker nodes (CEs and WNs). As soon as it is running, a job transfers all the files stored on the Storage Elements (SEs) via the Data Management System of the Grid (DMS) with the GridFTP protocol. During the job lifetime the status is retrieved from the user interface, and statistics are generated and collected to a remote server, which hosts a relational database and outputs these statistics through a web site. Once the job is finished, the outputs are stored back on the Grid Storage Elements via the Data Management System and the useful docking results are inserted directly from the Grid to a relational database where they can later be more easily queried and analyzed.

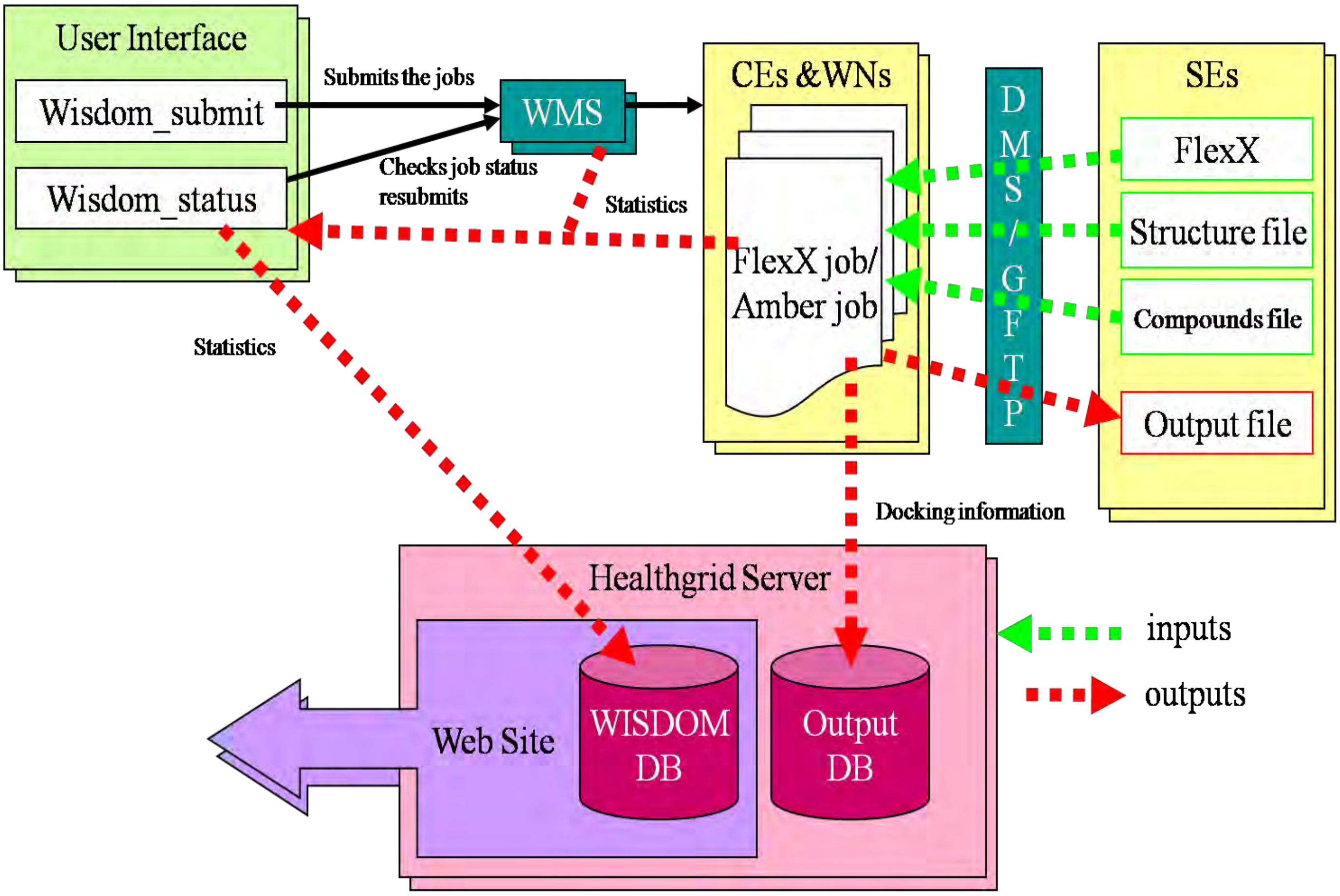


Figure 14: Schema of the WISDOM production environment utilized in WISDOM-II project.

3.3.3 Large-scale docking by using WISDOM environment

The deployment was performed on the previously listed infrastructures, and involved at least one manager to oversee the process on each of them. The three groups of targets (GST, *Plasmodium vivax* and falciparum DHFR) were docked against the entire ZINC database (4. 3 millions of compounds). The chemical compound database was split into 2,422 chunks of 1,800 compounds each. This is, because we wanted to have an approximated processing time ranging from 20 to 30 hours for each job (one docking process takes from 40s to 1min). The compound subsets have been stored on the involved Grid infrastructures. They were basically copied on a storage element (SE), and registered on the Grid data management system (DMS), and were also replicated on several locations whenever possible to improve fault-tolerance. We defined a WISDOM instance as, a target protein docked against the whole ZINC database, with a given parameter set. The Table 5 shows the various instances deployed on the different infrastructures.

A total number of 32 instances were deployed, corresponding to an overall workload of 77,504 jobs, and up to 140,000,000 docking operations. Of the total 32 instances, 29 were docked on EGEE, and 3 were run on AuverGrid, EELA and EuChinaGrid respectively.

As shown is Figure 14, the environment included a FlexLm server that provide the floating licenses for the FlexX commercial software. The FlexX software binaries were stored like all the inputs on the Grid storage elements (SE), and were installed on the fly on each worker node (WN) at the beginning of the job.

As the average duration of a job was around 20-30 hours, we submitted one instance per day, with a delay of 30 seconds between each submission. As one instance was submitted in about 20-30 hours, the submission process was quite continuous during the first month of deployment. The jobs were submitted to 15 Resource Brokers (the components of the Workload Management System) in a round-robin order. At the end of a job, the results were stored on the Grid storage elements, and directly into a relational database.

Distribution of the jobs

The repartition of the jobs on the different Grid federations is shown in Figure 15. It is showing also the contribution by the AuverGrid, EUChinaGrid and EELA infrastructures. Each of these three infrastructures ran one single instance which corresponds to 3% of the total 32 instances. The job repartition is quite similar to the previous deployments, but here the United Kingdom and Ireland federation (UKI) played an even bigger part. For instance, one of the British sites offered for quite a long period of time more than 1,000 free CPUs, which is half of the average used CPUs. Such a number of free resources available explain the repartition of the jobs. It is indeed a good view of the repartition of available resources during the deployment.

Target structures	*Number of instances deployed*
GST (A chain)	4 on EGEE
GST (B chain)	4 on EGEE
2BL9 (*P. vivax* wild type DHFR)	3 on EGEE, 1 on EELA
2BLC (*P. vivax* double mutant DHFR)	3 on EGEE, 1 on AuverGrid
Dm_vivax (*P. vivax* DHFR 2BLC minimized)	4 on EGEE
Wt_vivax (*P. vivax* DHFR 2BL9 minimized)	4 on EGEE
1J3K (*P. falciparum* Quadruple mutant DHFR)	4 on EGEE
1J3I (*P. falciparum* Wild type DHFR)	3 on EGEE, 1 on EuChinaGrid

Table 5: Instances deployed on the different infrastructures during the WISDOM-II data challenge One instance corresponds to one protein structure under one parameter condition and 4.3 million compounds.

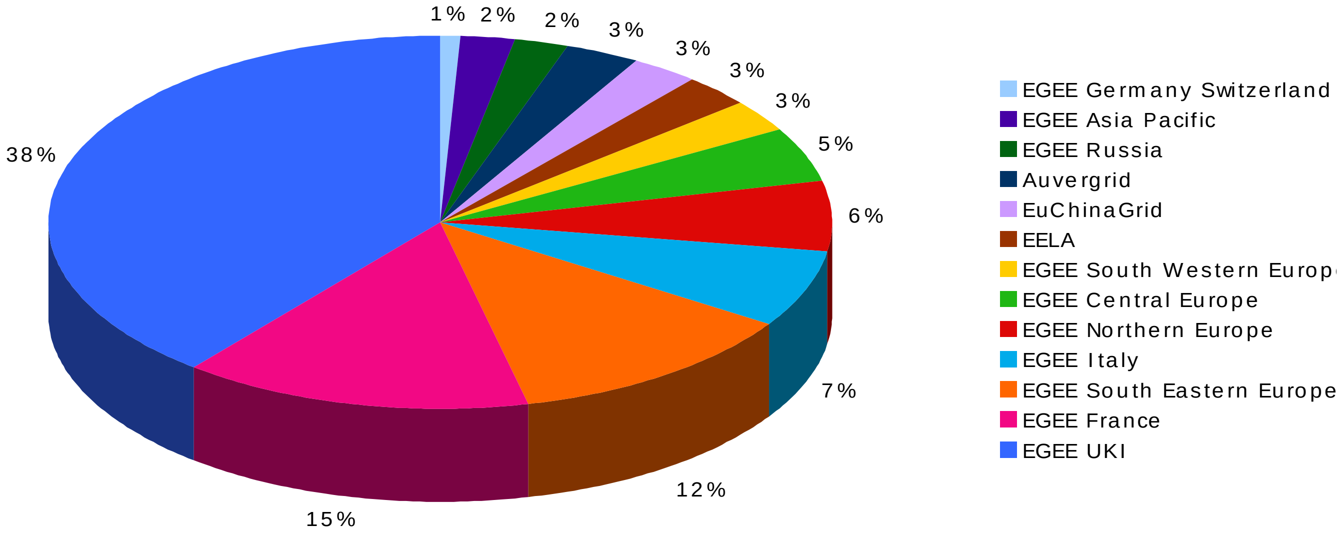


Figure 15: Distribution of jobs on the different Grid federations.
This repartition graph reveals the number of Grid infrastructures, and the countries involved in the large-scale screening (WISDOM-II). UKI federation contributed resources significantly, i.e., 38% of the total computing resources.

3.3.3.1 Results

The overall statistics of the deployment are shown in Table 6. The number of jobs mentioned in the first row of the Table 6, in fact corresponds to the number jobs that gave desired results. However, in reality, more jobs were actually submitted on the Grid, many of them were aborted and never really run on the Grid. When a job was done on the Grid, the WISDOM production environment checked the status file specifying the final result of the job: a job can be “Done” in the point of view of the worker node, without having produced the wanted results (docking results). In this specific case, the status of the job, which was stored on the Grid, was labelled as failed, and therefore, the environment has to resubmit the job again.

Number of Jobs	77,504
Total Number of completed dockings	156, 407,400
Estimated duration on 1 CPU	413 years
Duration of the experience	76 days
Average throughput	78,400 dockings/hour
Maximum number of loaded licences (concurrent running jobs)	5,000
Number of used computing elements	98
Average duration of a job	41 hours
Average crunching factor	1,986
Volume of output results	1,738 TB
Estimated distribution efficiency	39%
Estimated Grid success rate	49%
Estimated success rate after output checking	37%

Table 6: Overall statistics of the large-scale docking deployment (WISDOM-II).

In some cases, the environment failed at retrieving the status from the Grid, and thus considered implicitly the job has failed, even if the job has succeeded. It explains why some jobs ran several times, and why the final completed docking number is bigger than the useful desired dockings.

The average docking throughput is consistent with the crunching factor, which represents the average number of CPUs used simultaneously all along the data challenge. If we consider 80,000 dockings per hour for 2,000 CPUs (the crunching factor), it means 40 dockings for one CPU per hour, which is consistent with the empiric observation of one docking process lasting approximately 1 minute on a 3.06 MHz Intel Xeon processor. In fact, this is a little less than the estimation, because the actual duration of a single docking process on the Grid is a bit longer than the observed empiric duration (because of the overhead generated by the Grid). In the same logic, we can say that the instantaneous throughput peak would be obtained, when the max number of CPUs was used (i.e., 5,000), giving a throughput of approximately 200,000 dockings per hour.

The estimated Grid success rate is defined as the ratio between successful Grid jobs to that of the total of submitted jobs. The success rate after output checking will consider just the jobs that succeeded in producing the desired results, that’s why this score is lower. One can notice that these values are very small, but there are several explanations for this. In reality, at the beginning of the data challenge, the observed Grid success rate was about 80 to 90%, but it decreased constantly because of “non-performing” sites and the resource broker (RB) failures, mainly due to overload. Sometimes the available disk space was decreasing on some resources brokers, up to a point where some of the job data could not reach the computing element. In other cases, the sites were simply producing a lot of aborted job for an undetermined reason. The resource brokers failed again to balance the jobs, reasonably on the Computing Elements, and some of them ended up with more than 500 jobs in queue. At this point, the site administrator had no other choice, than kill all these jobs, producing in a single row more than 500 aborted jobs. Essentially, because of the automatic resubmission, this information should not be considered as way to evaluate the efficiency of the Grid, because the automatic resubmission guaranteed a successful job, and the aborted jobs are not staying on the Grid for long time consuming useful resources, simply because the majority of them were aborted before running. Moreover, if on an extreme, we decided to send all the jobs on a single working computing element, we could have achieved a 100% Grid success rate, but we would never have achieved a crunching factor of almost 2,000. So one must keep in mind that the Grid is a very dynamic system, and errors can occur at the last minute. Moreover, we ran these deployments during a transition period between several middleware components: the LCG was moving to gLite, and the ldap Virtual Organization system was moving to VOMS Virtual Organization system, with VOMS proxy extensions that expire after 24 hours, which could as well lead to a premature abortion of the jobs.

Issues related to docking deployment

As pointed out in the previous section, the scheduling efficiency of the Grid is still a major issue. The resource broker is still the main bottleneck, and even if used at high number (>15), it is always a source of trouble. Moreover, things get worse as load is increasing on the Grid. The « sink-hole » effects can result in sites overloading in a very short amount of time, and if not taken care immediately can lead to an impressive overhead caused by the long lasting waiting state of the jobs. One another reason was, that sometimes unreliable and incomplete information was provided by the information system, which does not publish the available slots and VO limitations, which would be mandatory to perform an efficient scheduling. Indeed, some sites may have several free CPUs announced, but if the maximum number of slots available is reached already, then the jobs submitted to this site will be queued. This deployment also shows that, it is not possible to do a naive blacklisting of the failing resources, for the simple fact that virtually all the Grid resources have produced aborted jobs.

Another issue was the ability to store and treat the data in a relational database. The machine hosting the database must have good performances; else, the number of queries coming from the Grid may overload the database management system significantly. In this deployment, we used a MySQL database, and planned to put all the produced results in the same table, but finally, we had to split this database in several ones (one per target), because the MySQL system did not scale and would not have been able to withstand the total number of records. The same comment goes with the storing of the jobs status. At one point, the throughput of arriving status collected was such, that the script that was supposed to treat them was generating CPU overloads on the machine, which lead to serious slowdowns.

All these elements clearly demonstrate that, even if the Grid can show very good result in comparison to simple architectures, it is still missing robustness and reliability. Nevertheless, performance wise, it can be improved.

3.3.4 Molecular dynamics on Grid

AuverGrid

Molecular dynamics deployments were achieved on AuverGrid [196], a French national Grid infrastructure. AuverGrid offers a computing power of more than 850 CPUs as well as 85TB disk space for data storage. The technology deployed on AuverGrid uses the gLite middleware, and is fully compatible with EGEE.

Deployment procedure

Unlike the deployment of molecular docking, which was performed on millions of compounds, as described in previous sections, molecular dynamics (MD) simulations were performed on the top scoring few thousands compounds that were resulting from these docking experiments. One of the main reasons to perform MD on limited number compounds was that MD simulations require much heavier computing time than docking. The MD procedure described here takes ~20 mins to simulate one compound (on a 3.06 Intel Xeon processor).

Four computing elements and two storage elements were used for the deployment of molecular dynamics simulation tools. After the simulations were finished, the result files were stored back on the storage element, and replicated twice on other locations for the backup. The molecular dynamics deployment was acheived in four steps:

a. Storage

The necessary files and executables for molecular dynamics deployment were pre-compiled Amber9 executables, protein structures, chemical compounds and the workflow script. In order to reduce the input-output operations, all the required input files and Amber executables were stored on the storage element (SE) of the Grid.

b. Testing

Test runs were performed before the large-scale deployment of Amber on Grid. This is done, in order to check, if there were any errors arising from the workflow script, and further to check the hardware influence on the ultimate results. However, when tested on small dataset (~100 compounds) on local machine and the Grid, the results were identical, thus, proved no hardware influence on the results.

c. Execution Procedure

The same deployment procedure and the production environment employed in docking were utilized here. Approximately 100 CPU were used (Clermont-ferrand computing element), during the execution process. As Grid generates overhead of several minutes, consequently, one should submit long enough jobs to neglect the effect of such an overhead. For this reason, it was decided to submit jobs lasting for approximately 20 hours (similar to docking jobs).

The first deployments were performed against 5000 best scoring docking poses of plasmepsin target. Jobs lasting ~20 hours were submitted, corresponding to 50 compounds in each subset (100 subsets were generated from 5000 compounds). Each subset with 50 compounds has been submitted on one worker node along with the Amber executables, target structures, chemical compound subsets, and the workflow script. After the simulations are finished, the final results: mmpbsa scores, and the three variations of protein-ligand complexes (initial docked complex, minimized complex and complex after molecular dynamics and re-minimization) are stored back on the storage element. In the similar way, 15000 docking poses of wild type Dihydrofolate reductase (wtDHFR), and 5000 docking conformations of Glutathione-S-transferase were deployed on the AuverGrid infrastructure.

d. Status

Job status on the Grid are frequently checked by using checkit.sh (a bash script, part of the WISDOM production environment), finally if the job is successful it gives a message "ok success"

3.3.4.1 Results

The total estimated CPU time, if the same simulations were to perform on one machine was expected to be 347 days. By using AuverGrid infrastructure, the simulation time was significantly brought down to 25 days. In total, 3 storage elements, 4 resource brokers, average 90 worker nodes were used in parallel for the deployment of 25,000 docking poses (5000 plasmepsin compounds, 15000 wild type DHFR compounds and 5000 GST compounds). Over all Grid statistics in the deployment of molecular dynamics simulations against 25000 compounds are displayed in Table 7.

Number of Jobs	500
Total Number of compounds simulated	25000
Estimated duration on 1 CPU	347 days
Duration of the experience	25 days
Maximum number concurrent running jobs	90
Number of used computing elements	1
Average duration of a job	20 hours

Table 7: Statistics of molecular dynamics simulations on Grid.
This Table demonstrates the significant gain in CPU time by using Grid computing. It demonstrates that if the experiment were performed on a single machine, it would have taken 347 days, which on Grid was done in 25 days.

Issues in molecular dynamics jobs submission

Amber is also commercial software with an integrated academic license designed in way that, it can be used on the cluster at an organization or research institute. In a normal installation, the Amber software is restricted to use on the local cluster at an institute. However, after negotiations, the University of California (the owner of the Amber software) allowed us to use of the software through the Grid on the clusters of the collaborative institutes, which possess the license rights. This, however, restricted us to deploy molecular dynamics simulations to only few computing elements.

Another observation was that due to unknown reasons, out of the 5000 compounds submitted against plasmepsin, ~100 compounds did not give the desired results (the MM-PBSA scores and MM-GBSA scores). Perhaps, this may be due to failure of resource broker while scheduling the jobs. The failed jobs were identified, and re-submitted again by black-listing the resource brokers, which were failing repeatedly.

3.4 Summary

In this chapter, the concept of e-Science and Grid computing is described in detail. Though computational Grids can be classified in different ways, depending upon the services offered by the Grid, hardware type and, underlying Grid technologies, this chapter provides two different types of classification with several example Grid projects; mostly in context to life science applications. This chapter also describes modern concepts such as Web services and Workflow systems.

Further, this chapter describes the deployment of large-scale molecular docking and molecular dynamics simulations on four different targets implicated in malaria. Several Grid infrastructures were used to achieve these deployments. However, resources from the EGEE Grid infrastructure was used extensively, this is because EGEE is one of the largest open source resource providers for performing research on neglected diseases. Other infrastructures that were utilized in the deployment of docking and molecular dynamics include EELA, EUChinaGrid and EUMedGrid and AuverGrid.

In context to molecular docking, the ZINC database, which consists of 4.3 million chemical compounds was screened against four different malarial target proteins under various receptor and docking software conditions. Screening of 4.3 millions compounds was achieved in 76 days on the Grid, which on a single machine would have taken 413 years. We have achieved an average docking throughput of 78,400 dockings. During the large-scale deployment, several issues were identified and several lessons were learnt. On the Grid side, the major issue identified was, repeated failure of resource brokers while scheduling the jobs. This led to overloading of the sites and further led to killing of jobs by the site manager.

At the end of the data challenge (Virtual screening at large-scale), the ultimate results have to be analyzed by biologists or chemists or bio-chemists, who sometimes possess no or little knowledge on Grid computing. One the other hand, handling the huge amount of data as flat files and analyzing them by using scripting languages or by any other means is quite challenging (especially, when the data analyzers are biologists). To overcome these issues, and to ease the result analysis of virtual screening data, MySQL databases are used.

Using the same WISDOM production environment that was used for docking deployment, molecular dynamics simulations were performed. In less than 25 days, MD simulations were performed on 25,000 best docking conformations, which on single CPU would have taken 347 days. Similar issues that were identified during docking deployment (resource broker failure) were also observed in molecular dynamics simulations deployment. To the best of our knowledge, this is the first time a molecular dynamics application is deployed on the Grid in embarrassingly parallel way.

Though several potential issues were identified during the deployment, computational Grids reduced the overall time required for screening and simulating thousands of chemical compounds. In summary, we can conclude from this chapter that, the computational Grids are a part of e-Science paradigm that opens new means to perform biomedical research and make possible to perform large-scale experiments such as molecular docking and molecular dynamics simulations easily.

4 Chapter 4. Discovery of plasmepsin inhibitors by large-scale virtual screening

This chapter reports the complete set up of a large-scale virtual screening against plasmepsin by molecular docking application with a special focus on modeling and evaluation aspects. This chapter is organized as follows:

Firstly, section 4.1 describes the role of haemoglobin degradation in plasmodium survival and the role of plasmepsins in the metabolism of haemoglobin as well as its suitability as a drug target in the current thesis. As the structural features, crystallographic information and active site information of the target protein plays a significant role before and after the virtual screening, it is given in section 4.1. In section 4.2 and 4.3, the material utilized for the screening, the overall architecture, algorithm, scoring function of FlexX and AutoDock software and the chemical compound database, the Chembridge database are discussed in detail. In section 4.4, the experimental setup is demonstrated, with the focus on requirements for different parameter settings on the side of the target and on the side of the docking software. In section 4.5, the need for storing different types of results for each docking and novel strategies in analyzing the results are explained. Later in section 4.5, the results, and the modeling aspects of the top scoring compounds are discussed. Finally, in section 4.6, the summary of this chapter along with potential issues and probable solutions is given.[2]

4.1 Haemoglobin degradation

Hemoglobin degradation is an essential process of the exo-erythrocytic cycle of Plasmodium, and is the key pathways in plasmodium survival. For this reason, most of the antimalarial drugs are aimed at disrupting the hemoglobin degradation pathway [198]. The Plasmodium parasite metabolizes most of the host cell hemoglobin inside the erythrocyte, during different intra-erythrocyte stages: ring stage, trophozoite stage and schizont stage. The metabolic activity varies between the various stages and is more pronounced during the trophozoite stage [198]. The hemoglobin degradation takes place in a specialized organelle called food vacuole. The majority of hemoglobin from the erythrocyte cytosol, is targeted to the Plasmodium food vacuole through a cytostomal system. Once in the food vacuole, when hemoglobin experiences an acidic pH of 5 to 5.5, proteolysis of hemoglobin occurs in a

This chapter is based on Kasam, V., et al, Design of Plasmepsin Inhibitors: A Virtual High Throughput Screening Approach On The EGEE Grid, J. Chem. Inf. Model. 2007, 47, 1818-1828.

seemingly ordered pathway. This results in the generation of small peptide fragments of hemoglobin [199, 200]. These peptides are transported out of the food vacuole to the cytosol, where they undergo further degradation to yield individual amino acids. There are two unwanted byproducts of hemoglobin degradation. First by product is heme, which is highly reactive and toxic to the cell. Detoxification is achieved by sequestration of heme in the form of polymers called hemozoin. This process of heme polymerization is unique to Plasmodium, and for this reason, inhibition of this process has been a major area of investigation. There are several proteases involved in heme metabolism, but the present work focuses on plasmepsin family of proteins. The role of plasmepsins in heme metabolism is described in following sections.

4.1.1 Plasmepsins

Plasmepsins (Plm) are involved in the hemoglobin degradation inside the food vacuole during the erythrocytic phase of the life cycle. There are ten different isoforms of this protein species, and ten genes are encoding them in *Plasmodium falciparum* (Plm I, II, III, IV, V, VI, VII, IX, X and HAP). Other human Plasmodium species contain lesser number of plasmepsin isoforms than that of *Plasmodium falciparum* [201]. Expression of Plm I, II, IV, V, IX, X and HAP occurs during the erythrocytic cycle, and expression of Plm VI, VII, VIII, occurs in exo-erythrocytic cycle [202].

Mode of action

The complete hemoglobin degradation inside the food vacuole of the Plasmodium is illustrated in Figure 16. The two homologous plasmepsins, I and II, are responsible for the initial attack on the hemoglobin alpha chain between the residues Phe33 and Leu34, in the hinge region [203]. This region is highly conserved and responsible for the stability of the hemoglobin tetramer. Upon cleavage, heme (ferrous +2) is released, which is toxic to the parasite.It is further oxidized to hematin (ferric +3), but even after oxidization, hematin is toxic to parasite. Finally, the hematin is polymerized to hemozoin (the so-called malarial pigment). The globulin part of the hemoglobin is further metabolized by carboxypeptidases to amino acids. The parasite relies, and feeds on these amino acids for its survival. Both plasmepsin I and II are capable of making an initial cleavage in the hemoglobin, and other plasmepsin isoforms make several other cleavages after the initial attack [204].

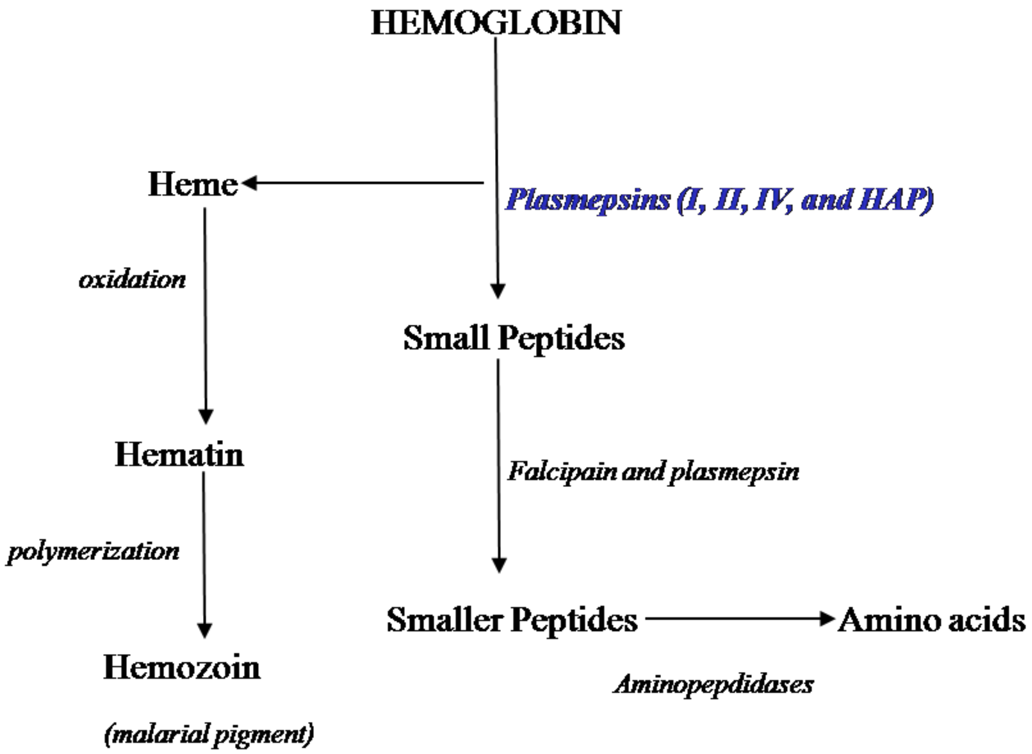


Figure 16: Pictorial representation of hemoglobin degradation [204].
The Figure demonstrates that the hemoglobin degradation starts with plasmepsin members of the family. Plasmepsin degrades hemoglobin into heme (indicated on the left hand side) and small peptides. Further heme is oxidized into hematin and hemazoin. The small peptides are further metabolized by other proteases and aminopeptidases to amino acids (indicated on right hand side).

Homology and selectivity

Target selectivity is the major barrier for the identification and development of novel inhibitors, especially when the drug development involves pathogenic system [205]. Selectivity (selection of the target protein involved in the disease over its nearest human (desired) protein) is the one of the key aspect to be considered before setting up a drug discovery campaign. Hence, plasmepsin sequence similarity is checked at two levels; firstly, similarity is checked between different isoforms of plasmepsin family, and secondly to its nearest human aspartic protease, cathepsin D. High levels of sequence homology are observed between the different plasmepsin subtypes, Plm I, II, IV, and HAP, which lie in the cluster of the same gene. The sequence similarity at the binding site region of Plm II to that of Plm I, IV, and HAP are 84%, 68%, and 39% respectively. This observation enable newly found inhibitors effective on family of proteins rather than single subtype.

The selectivity of a drug between the parasite and the closely related human aspartic proteases is one of the important considerations for the development of new compounds against plasmepsins. The closest human aspartic protease, cathepsin D has fortunately only 35%

overall sequence similarity to Plm II [206, 207]. The low sequence similarity of plasmepsin over the human cathepsin D suggests that the newly found inhibitor will be effective only on the pathogen system without affecting human system. In contrast, if the sequence homology between the plasmepsin and cathepsin D (or any other human protein) were high, the main concerns would have been toxicity and reduced concentrations of the drug reaching the target protein of the pathogen system.

4.1.2 Structural information of plasmepsins

Twelve different X-ray structures of Plm II are presently available in the Brookhaven protein database (www.pdb.org). The current study is preceded with four structures, three of them are monomers (1LEE, 1LF2, 1LF3), and one protein is a dimer (1LS5, which presumably is a crystallographic artifact since the biologically active form is a monomer). All the proteins are co-crystallized with different types of inhibitors (peptidic and non-peptidic inhibitors).

Target ID	Crystallization method	Resolution Å	Ligand	Nature of the ligand	Number of Monomers
1LEE	X-Ray	1.9	R36	Non-Peptidic	1
1LF2	X-Ray	1.8	R37	Non-Peptidic	1
1LF3	X-Ray	2.7	EH5	Non-Peptidic	1
1LS5	X-Ray	2.8	IHN48	Peptidic	2

Table 8: Represents the crystallographic features of plasmepsin targets utilized in the current thesis.

Active site description of 1LEE and 1LF2

Each crystal represents one monomer per asymmetric unit. Both inhibitors have a Phe-Leu core and incorporate tetrahedral transition state mimetic, hydroxypropylamine. The inhibitor R36 (1LEE) possesses a 2,6-dimethylphenyloxyacetyl group at the P2 position and 3-aminobenzamide at the P2' position, while R37 (1LF2), posess the same P2 group but 4-aminobenzamide in the P2' position. Figure 17 illustrates the R36 (1LEE ligand) and R37 (1LF2 ligand) in the active sites of their respective proteins. These complexes reveal key conserved hydrogen bonds between the inhibitor and the binding-cavity residues, notably with the flap residues Val78 and Ser79, the catalytic dyad Asp34 and Asp214 and the residues Ser218 and Gly36 that are in proximity to the catalytic dyad. The structures also show unexpected conformational variability of the binding cavity of plasmepsin II, and may reflect the mode of binding of the hemoglobin alpha-chain for cleavage. This confirms that the target structures are flexible at their binding site regions [208].

The structures of the R36 (1LEE) and R37 (1LF2) complexes are virtually identical except for the difference in the position of the amino groups in the P2' inhibitors (Highlighted in circles in Figure 17). The RMS deviation after superimposing the main-chain atoms of the two models is 0.201 Å. (This value corresponds to all the atoms in the proteins including ligands). There is a negligible difference in the inhibition-constant values of both R36 (1LEE) and R37 (1LF2) against PLM II, with inhibition-constant values of 18 nM for R36 (1LEE) and 30 nM rs370 (1LF2) [208]. As the resolutions for 1LF3 2.7Å and 1LS5 2.8Å are sub optimal, the structural details are not discussed in detail.

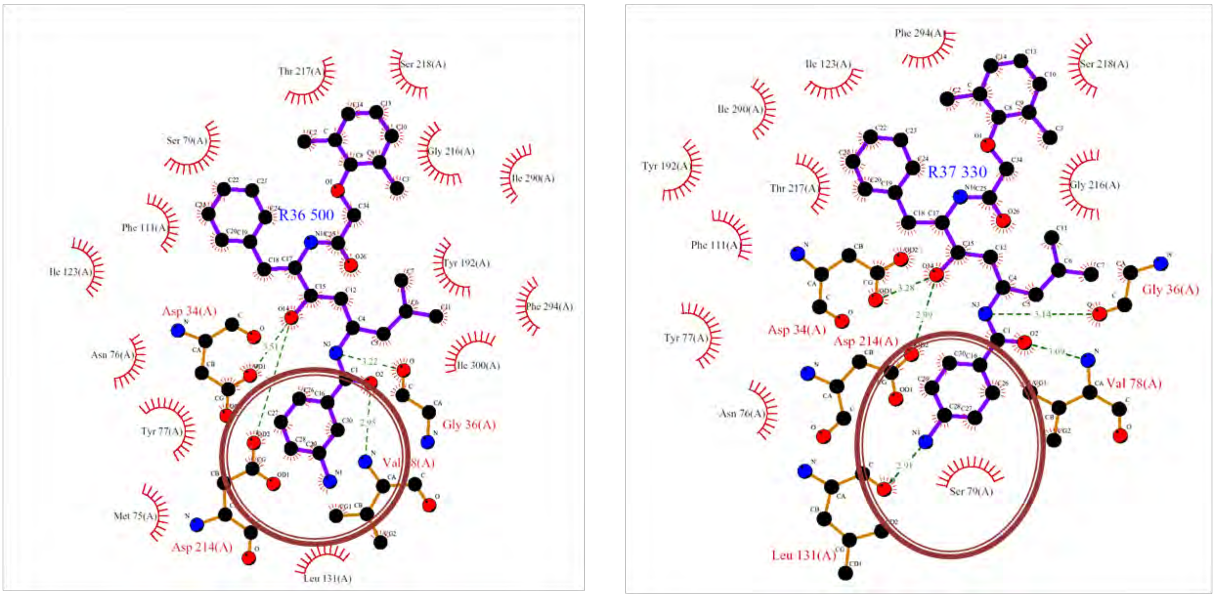


Figure 17: Ligand plots of target structures 1LEE (left) and 1LF2 (right).
The ligands are represented in ball and stick model in CPK colour. Hydrogen bonding between the ligand and the active site residues are indicated in green colour dotted lines and hydrophobic environment is indicated in crecent shape. The difference between the two ligands are highlighted in red circles. The plots are obtained from www.pdb.org.

The ligand plots of the target structures 1LEE and 1LF2 (pdb id) are shown in Figure 17, these plots are obtained from Brookhaven protein database (www.pdb.org). It is clear from Figure 17, that, in both the structures the inhibitors R36 and R37 (O14 atom of the inhibitor) form hydrogen bonds with catalytic residues (ASP34 and ASP214) of the protein. The interactions to the catalytic dyad are highly conserved throughout the plasmepsin family. In addition to the hydrogen bonding, other key interactions and hydrophobic environment around the ligand are clearly illustrated in Figure 17. These plots not only give an idea on the binding mode of the ligand, but also help in defining the active site, while preparing the target structures for the docking experiments. In the Figure 17, hydrogen bonds are indicated in green dotted lines, hydrophobic environment is represented in red color semi circles. The ligands (R36 and R37) are shown in ball and sticks form. The catalytic residues ASP34, ASP214, and flap residues Val78 and Gly36 are also shown in ball and sticks style.

Plasmepsin as drug target

For a protein to be considered as a validated target for protein based virtual screening, it has to meet the following criteria

a. It has to play a key role in the survival of the organism
b. Availability of sound crystallographic data.

Based on the discussions from the above sections, it is clear that plasmepsin plays a key role in the Plasmodium survival and posses of good quality X-ray crystal data. This makes members of plasmepsin family, ideal targets in anti-malaria therapy, and in rational drug design.

Target preparation

A good structure (X-ray model, which is crystallized at low resolution, typically <2Å), and careful target preparation greatly affects the results obtained in the virtual screening process. The 3D co-ordinates of all the proteins used in the present study are obtained from the Brookhaven protein database. (www.pdb.org).

In the first step, in order to transfer and compare binding modes between different receptor structures, all the water molecules and co-crystallized ligands are removed. In the next step, all the plasmepsin structures are superimposed on 1LEE (PDB ID). 1LEE serves as a reference template. Figure 18 illustrates the superimposition of all the five-plasmepsin structures. The active site comprises all atoms within 6.5Å of the co-crystallized ligands as well as residues of known relevance (see ligand plots). The charges of the ionizable groups are chosen to be consistent with acidic conditions (pH 5). Hemoglobin degradation by plasmepsin takes place inside the acidic food vacuole, where the pH conditions are acidic pH (pH 5); to be consistent with biological environment, the acidic pH conditions are chosen. The side chains of lysine and arginine residues are protonated, as well as the side chain of histidine is protonated (for comparability reasons, since AutoDock regards all histidine residues as protonated). The carboxylic groups of glutamic acid and aspartic acid are deprotonated.

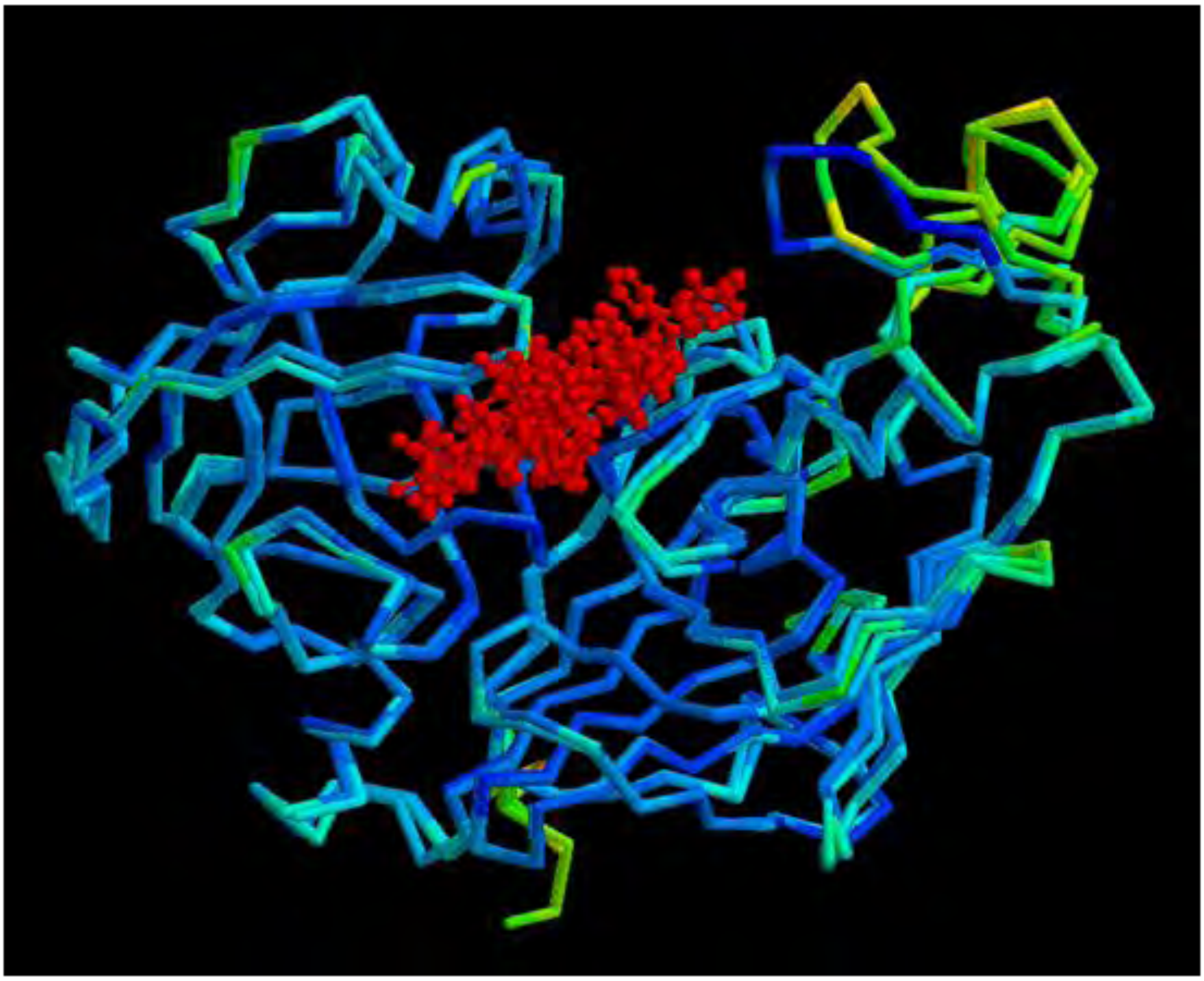


Figure 18: Screen shot of five plasmepsin structures superimposed.
Variations exist in the loop regions and are highlighted in yellow color. The co-crystallized ligands are are represented as balls and sticks in Red color. The picture has been generated using Rasmol.

4.2 Compound database selection

Compound libraries used in the virtual screening experiments should be filtered first, to remove unsuitable compounds that would not reach and pass the clinical trials due to undesired and toxic properties. Usually, compounds are filtered based on their chemical descriptors. Additionally, similarity searching methods or pharmacophore based screening methods or other ligand-based methods are often used, to increase hit rate or to reduce size of compound database, prior to molecular docking. However, there is a possibility that the filtering methods such as similarity searching and pharmacophore-searching methods could eliminate the possible lead compounds before molecular docking.

Furthermore, good binding affinity data for a series of ligands with similar mode of action (or mode of binding) is required to build a good pharmacophore model and that in most cases, such high quality data are missing. On the other hand, similarity searches are “searching under light” and in essence, they are mining approaches in “known territory.”

As a compromise between the removals of unwanted compounds, and not losing any lead compounds prior to screening, a very popular method to evaluate the drug likeness of a candidate structure is the so-called Lipinski “Rule-of-five”, is utilized in the current thesis. The Compound library used was obtained from the ZINC database [209, 210]. The ZINC database is a collection of 3.3 million chemical compounds (in 2005) from different vendors. The ZINC library was chosen because, it is an open source database, the structures have already been filtered according to the Lipinski rules and the data are available in different file formats (Sybyl mol2 format, sdf and smiles). Therefore, basically, ZINC provides virtual compounds ready for virtual screening. One million compounds were downloaded from the ZINC database, which includes 500,000 compounds from ChemBridge (vendor) [211] and, 500,000 additional drugs like compounds from various other vendors.

As AutoDock requires the pdbqs file format, all the compounds were first converted into pdbqt format using ADT tools. For, FlexX there was no such formatting was required. The Figure 19 displays the range of descriptor values of the chemical compound database. It is extremely important to note that all the compounds are in accordance with the “Lipinski rule of five”.

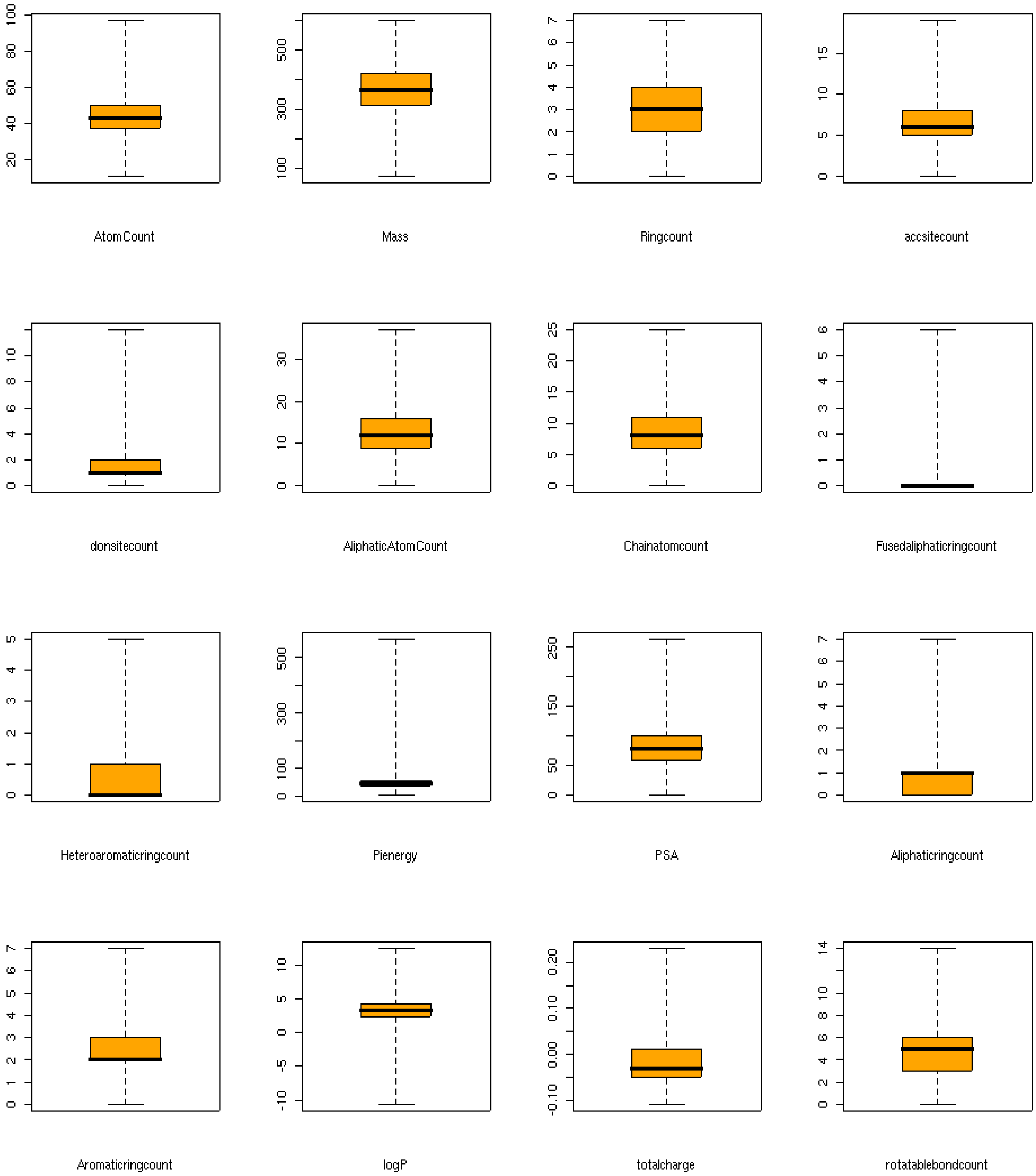


Figure 19: Illustrates descriptor values of Chembridge chemical compound database. Along with the descriptors suggested in “Lipinski rule of five”, several other important descriptor values were calculated and are illustrated. All most all the compounds display compounds possess acceptable descriptor value.

4.3 Docking software

Hit identification can be improved by enabling the virtual screening process by utilizing several docking tools [98]. Hence, for the virtual screening, two different docking software are used: FlexX 2.0 [78, 84] and AutoDock 3.05 [93, 212].

Docking and Scoring in AutoDock

There are three different types of algorithms in AutoDock [93, 212]: simulated annealing, Lamarkian genetic algorithm and genetic algorithm. In the current thesis, the genetic algorithm (GA) with local search is used.

General steps in the genetic Algorithm

a. Start with a random population (50-200).
b. Perform Crossover (Sex, two parents -> 2 children) and Mutation (one individual gives 1 mutant child).
c. Compute fitness of each individual.
d. Proportional selection and Elitism.
e. New Generation begins if total energy evaluations maximum generations are not reached.

AutoDock parameter sets used in the current thesis are:

1. GALS (Genetic Algorithm Local Search with Solis-Wets (SW))
2. GALS (Genetic Algorithm Local Search with pseudo Solis-Wets (pSW))

Docking and Scoring in FlexX

FlexX is an extremely fast, robust and highly configurable computer program for predicting protein-ligand interactions. Standard parameter settings are used except for two cases ("Place particles" [213] and "Maximum overlap volume" [84]). These two parameters were subject to deliberate variation with FlexX are given in Table 9.

Parameter sets	***Place particles***	***Maximum overlap volume***
Parameter set 1	Yes	2.5
Parameter set 2	Yes	5
Parameter set 3	No	2.5
Parameter set 3	No	5

Table 9: The parameter sets used during the FlexX data challenge.
Place particles and Max overlap volume are the two variations used in FlexX software.

Place particles

This is a special feature of FlexX, where the place particle algorithm places virtual water by itself. If it equals 1 (during the FlexX run), FlexX places spherical objects, called particles, into the suitable positions of the active site. Particles can then mediate interactions between the ligand and the protein. The main application is the modeling of discrete water molecules located between the protein and the ligand. If it equals to 0, FlexX does not add any water molecules by itself [213].

Maximum overlap volume

This is a clash test between the protein and ligand atoms. The condition for a clash is that a protein and a ligand atom exhibit an overlapping volume greater than the maximum allowed value. Hydrogen atoms not considered in the overlap tests. An easy way to switch off the overlap test is to set this parameter to a very high value.

The default value set in FlexX software is 2.5 Å (Reasonable range: 0.0-100.0 Å).

Algorithm

FlexX uses an incremental construction algorithm [78] and Boehm's empirical scoring function [101, 42]. During the incremental construction, rigid portions of the ligand are docked first, followed by the flexible portions. The incremental construction algorithm samples the conformational space of the ligand and uses a hierarchical system for placing the flexible pieces of the ligand. The anchor portion of the ligand, or the base is selected first (called base selection), and is placed in such a way that the interactions between the fragment and the protein are maximized (called base placement). Many alternatives for the placement of the flexible portions of the ligand, starting with those nearest the base, are considered and only those with favorable energies are considered for sequential rounds during which additional flexible portions are added.

Scoring Function

The scoring in FlexX is based on the empirical Boehm scoring function [101, 42]. Both the hydrophobic and hydrogen bonding contribute to the final score of a particular protein-ligand interaction. However, the hydrogen bonding (hydrogen acceptor and donor) are weighted higher than interactions resulting from hydrophobic interactions, this is one reason, why ligands that are capable of making more hydrogen bonds scores better than ligands making hydrophobic interactions. More details about the FlexX algorithm and scoring function are provided in [78, 214].

Interaction types

The common interaction types possible between the protein and ligand, and their corresponding energy contribution in FlexX are given in Table 10. From the Table 10, it is evident that the energy contribution from hydrogen bonding (4.7kJ/mol) is significantly higher than the energy contribution of hydrophobic interactions (0.7 kJ/mol). This is one reason why ligands with more hydrogen bonds (either acceptor or donor) obtain good scores and further rank higher [214]. Interaction types are in the following order level 3>level 2>level 1.

Interaction types	Interaction distance (A)	ΔG neutral kJ/mol	ΔG ionic kJ/mol	Level
H-acceptor/ H-donor	1.9	4.7	8.3	3
Metal acceptor/ metal	2	4.7	8.3	3
Aromatic ring atom, methyl, amide/aromatic ring center	4.5	0.7	-	2
Aliphatic and aromatic carbon atoms, Sulfur	4.5	-	-	1

Table 10: Interaction types of FlexX and their corresponding energy contributions.
The values in the Table are adapted from [214].

4.4 Virtual docking process

4.4.1 Re-docking, cross docking and docking under different parameter sets

Direct docking and re-docking experiments are performed between the target structures and their respective co-crystallized ligands on different parameter sets for FlexX. Re-docking can be defined as the removal of the co-crystallized ligand (inhibitor or substrate), and then using a specific parameter set, dock this compound back into the active site of its target protein. Re-docking is done, to validate the program’s ability to dock novel compounds into the active site of the protein. These experiments serve as positive controls before the large-scale docking is done, furthermore, it aids in defining the active site and other simulation conditions. Validation of the docking pose is done by comparing the interaction information between the re-docking pose to the ligand plot information obtained from the Brookhaven protein database

PDB. The lower the RMSD value, and the more similar the docking poses to the co-crystallized ligand, the better are the re-docking results. Ligand plots illustrate the binding mode of the co-crystallized ligand within the active site of the receptor, and further describe the atom-to-atom interaction between the co-crystallized ligand and its respective receptor. Ligand plots of 1LEE and 1LF2 are illustrated in Figure 17.

Re-docking

Despite of the large ligand size (>15 ligand components and >12 rotatable bonds), the RMSD values in re-docking experiments for 1LEE and 1LF2 were convincing, about 2.5Å for the top ranking solutions. The scores and RMSD values of the re-docking are shown in Table 11. Figure 21 illustrates the binding mode of the ligand (R36) in re-docking experiment with target protein 1LEE. The RMSD values for 1LF3 and 1LS5 are marginal and >3Å in all the parameter sets, see Table 11. This may be due to the suboptimal resolution of the X-ray crystal structure, very large ligand size, numerous ligand components (co-crystalized ligand of 1LF3 has about 24 fragments) and consequently high number of rotatable bonds. In the re-docking experiment for llf3, we observed a rotation of a ligand (flip), while maintaining the essential contacts to the catalytic dyad, this may be due to particular crystallization conditions have such an influence, thus the calculated position is not necessarily wrong. The resulting binding modes for 1LEE and 1LF2 were well in concordance with the crystal structure. The binding mode of the best ranked solution displayed good interactions with catalytic residues of the targets (Asp214, Asp34) and also with other residues of relevance. This comparison revealed that the ligand in the protein structures 1LEE and 1LF2 has found all the significant interactions responsible for the activity of the protein. From parameter set 1, it became also clear that direct docking (targets without any crystal water) performed well both in terms of scoring and RMS deviations. From interaction information, it was observed that protein structure 1LEE without any crystal water molecules formed hydrogen bonds with both the catalytic residues Asp214 and Asp34, as well as with flap residues Val78 and Ser218. Surprisingly, the protein structure 1LEE with crystal water molecules, all ligands failed to form interactions with the key residues.

Target	Ligand	Total Score	RMS-Value	Total Score	RMS-Value	Total Score	RMS-Value	Total Score	RMS-Value
		1		**2**		**3**		**4**	
1LEE	1LEE (R36)	-21.128	2.34	-15.348	9.66	-28.075	9.82	-25.914	2.09
1LEE_h2	1LEE (R36)	-20.079	8.51	-14.806	9.51	-25.959	2.09	-25.959	2.09
1LEE_h3	1LEE (R36)	-26.197	9.80	-19.664	8.92	-26.431	9.81	-26.039	2.09
1LF2	1LF2 (R37)	-24.319	4.93	-23.401	3.26	-22.134	9.86	-24.672	9.37
1LF2_h	1LF2 (R37)	-19.563	10.03	-27.984	4.81	-22.962	2.77	-24.672	9.37
1LF3	1LF3 (E58)	-20.928	8.97	-19.461	13.78	-13.941	7.60	-16.921	12.46
1LS5_a	1LS5 pepstatin A	-23.219	10.87	-22.35	11.72	-27.677	3.22	-33.698	3.13
1LS5_b	1LS5 pepstatin A	-23.105	10.52	-20.708	12.20	-30.313	4.92	-24.86	11.71

Table 11: Illustrates docking scores and RMSD values for the best ranking solutions under four different parameter sets.

The Table displays RMSD values in direct docking and re-docking experiments obtained for four different parameter sets with FlexX. 1, 2, 3, 4 corresponds to parameter set 1, 2, 3, 4 respectively. Units for Total score and RMS value are kJ/mol and Angstroms respectively.

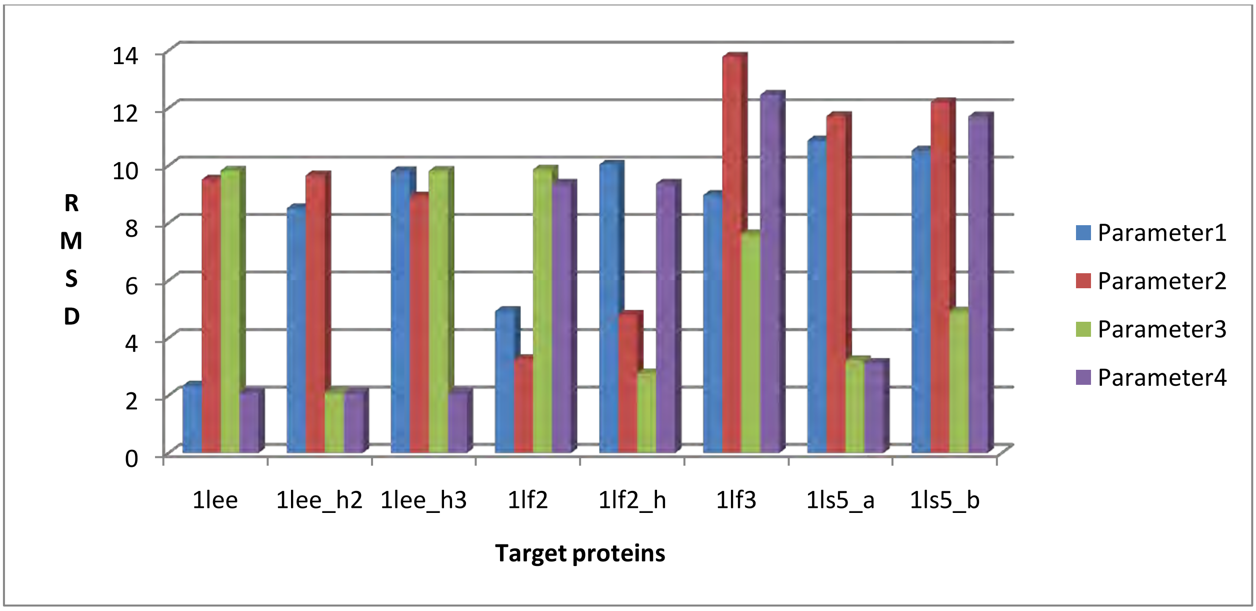
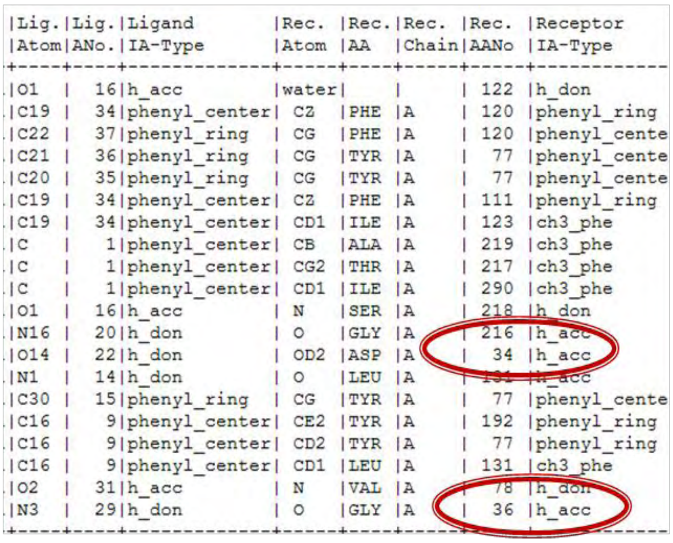


Figure 20: Illustrates the RMSD values in re-docking experiments under different parameters. The RMSD values are displayed on the Y-axis, and on the X-axis are the protein structures. The protein structure 1lee and 1lf2 were considered in several structural forms based on the inclusion of water molecules.

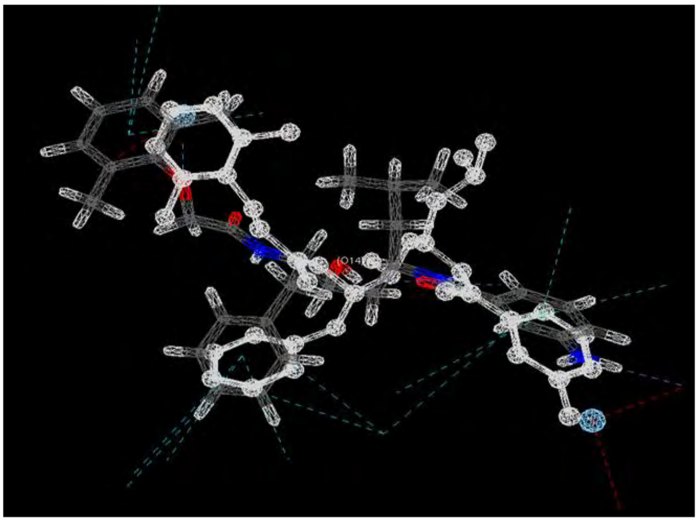
Effect of docking under parameter sets

Selection of the parameters is often challenging and parameters can be set only after preliminary testing on a particular dataset. Since every target has a unique response to the docking software parameters set, there is no generic solution to agree on a particular parameter sets. The success of the docking methods significantly depends on the parameter sets under which docking is performed. Initial optimization experiments were performed, to arrive at a parameter set which gave best results on known inhibitors, and in re-docking experiments for a particular target. Figure 21 and 22 and represents the results of re-docking of R36 ligand into its protein structure 1lEE. Table 12 lists hydrogen bonding of R36 with conserved residues and RMSD values under different parameter sets. There is no significant difference in docking score between the different parameter sets is noticed. This is due the fact that, the anchor fragment selected during the FlexX docking is always the same, hence similar scores are observed. But, different parameter sets produced different binding modes in the re-docking experiment. The RMSD value varied significantly (see first row of Table 11 and first histogram of Figure 20). Visualization of binding poses revealed that there is noteworthy variation in the binding poses in all the parameter sets. Especially, large variations were observed with parameter set 3 and 4 (See Figure 22). In parameter set 3, while maintaining hydrogen bonding to essential amino acids of the protein, the binding mode of the docking pose was completely flipped, this is one of the reason why the RMS deviation was high (>9) in parameter set 3. Figure 22 displays the re-docking pose in parameter set 3, the part of the ligand flipped is highlighted in circles. In paramter set 4, the docking pose is not flipped as in parameter set 3, but large deviations are observed at P2 position of the ligand (see Figure 22, variations at P2 positions are highlighted in circle). As large variations on the docking poses were noticed, we decided to perform large-scale virtual screening on all the parameter sets.

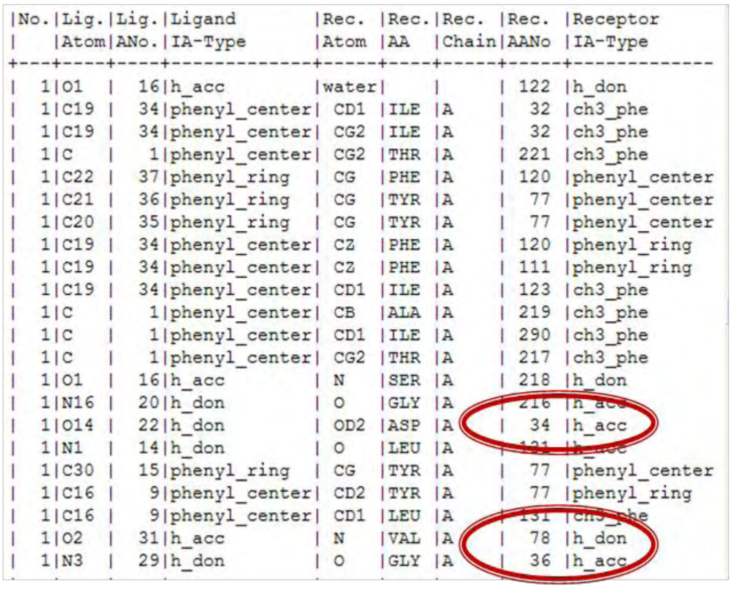
Parameter set 1



```
|Lig.|Lig.|Ligand        |Rec. |Rec.|Rec. |Rec. |Receptor
|Atom|ANo.|IA-Type       |Atom |AA  |Chain|AANo |IA-Type
+----+----+-------------+-----+----+-----+-----+-----------
.|O1  |  16|h_acc        |water|    |     | 122 |h_don
.|C19 |  34|phenyl_center| CZ  |PHE |A    | 120 |phenyl_ring
.|C22 |  37|phenyl_ring  | CG  |PHE |A    | 120 |phenyl_cente
.|C21 |  36|phenyl_ring  | CG  |TYR |A    |  77 |phenyl_cente
.|C20 |  35|phenyl_ring  | CG  |TYR |A    |  77 |phenyl_cente
.|C19 |  34|phenyl_center| CZ  |PHE |A    | 111 |phenyl_ring
.|C19 |  34|phenyl_center| CD1 |ILE |A    | 123 |ch3_phe
.|C   |   1|phenyl_center| CB  |ALA |A    | 219 |ch3_phe
.|C   |   1|phenyl_center| CG2 |THR |A    | 217 |ch3_phe
.|C   |   1|phenyl_center| CD1 |ILE |A    | 290 |ch3_phe
.|O1  |  16|h_acc        | N   |SER |A    | 218 |h_don
.|N16 |  20|h_don        | O   |GLY |A    | 216 |h_acc
.|O14 |  22|h_don        | OD2 |ASP |A    |  34 |h_acc
.|N1  |  14|h_don        | O   |LEU |A    | 131 |h_acc
.|C30 |  15|phenyl_ring  | CG  |TYR |A    |  77 |phenyl_cente
.|C16 |   9|phenyl_center| CE2 |TYR |A    | 192 |phenyl_ring
.|C16 |   9|phenyl_center| CD2 |TYR |A    |  77 |phenyl_ring
.|C16 |   9|phenyl_center| CD1 |LEU |A    | 131 |ch3_phe
.|O2  |  31|h_acc        | N   |VAL |A    |  78 |h_don
.|N3  |  29|h_don        | O   |GLY |A    |  36 |h_acc
+----+----+-------------+-----+----+-----+-----+-----------
```



Parameter set 2



```
|No.|Lig.|Lig.|Ligand        |Rec. |Rec.|Rec. |Rec. |Receptor
|   |Atom|ANo.|IA-Type       |Atom |AA  |Chain|AANo |IA-Type
+---+----+----+-------------+-----+----+-----+-----+------------
|  1|O1  |  16|h_acc        |water|    |     | 122 |h_don
|  1|C19 |  34|phenyl_center| CD1 |ILE |A    |  32 |ch3_phe
|  1|C19 |  34|phenyl_center| CG2 |ILE |A    |  32 |ch3_phe
|  1|C   |   1|phenyl_center| CG2 |THR |A    | 221 |ch3_phe
|  1|C22 |  37|phenyl_ring  | CG  |PHE |A    | 120 |phenyl_center
|  1|C21 |  36|phenyl_ring  | CG  |TYR |A    |  77 |phenyl_center
|  1|C20 |  35|phenyl_ring  | CG  |TYR |A    |  77 |phenyl_center
|  1|C19 |  34|phenyl_center| CZ  |PHE |A    | 120 |phenyl_ring
|  1|C19 |  34|phenyl_center| CZ  |PHE |A    | 111 |phenyl_ring
|  1|C19 |  34|phenyl_center| CD1 |ILE |A    | 123 |ch3_phe
|  1|C   |   1|phenyl_center| CB  |ALA |A    | 219 |ch3_phe
|  1|C   |   1|phenyl_center| CD1 |ILE |A    | 290 |ch3_phe
|  1|C   |   1|phenyl_center| CG2 |THR |A    | 217 |ch3_phe
|  1|O1  |  16|h_acc        | N   |SER |A    | 218 |h_don
|  1|N16 |  20|h_don        | O   |GLY |A    | 216 |h_acc
|  1|O14 |  22|h_don        | OD2 |ASP |A    |  34 |h_acc
|  1|N1  |  14|h_don        | O   |LEU |A    | 131 |h_acc
|  1|C30 |  15|phenyl_ring  | CG  |TYR |A    |  77 |phenyl_center
|  1|C16 |   9|phenyl_center| CD2 |TYR |A    |  77 |phenyl_ring
|  1|C16 |   9|phenyl_center| CD1 |LEU |A    | 131 |ch3_phe
|  1|O2  |  31|h_acc        | N   |VAL |A    |  78 |h_don
|  1|N3  |  29|h_don        | O   |GLY |A    |  36 |h_acc
```

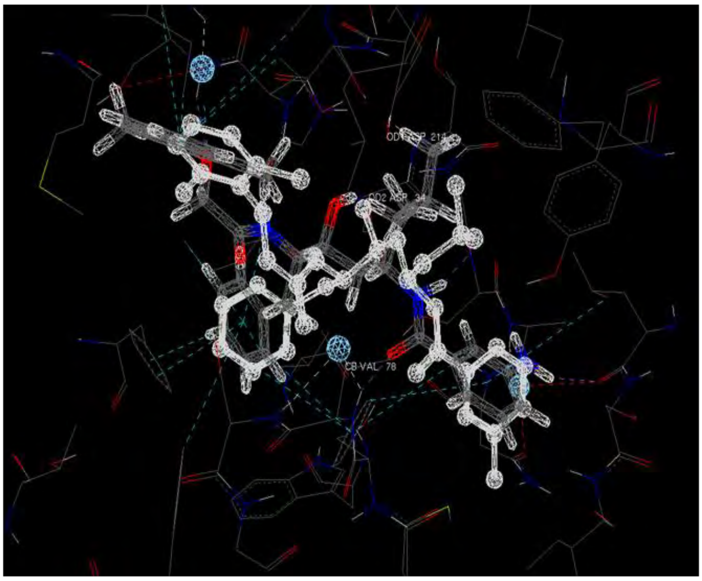
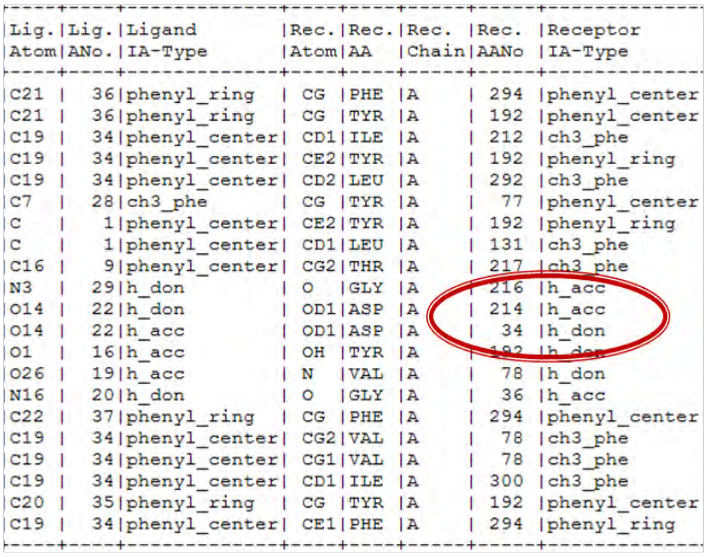


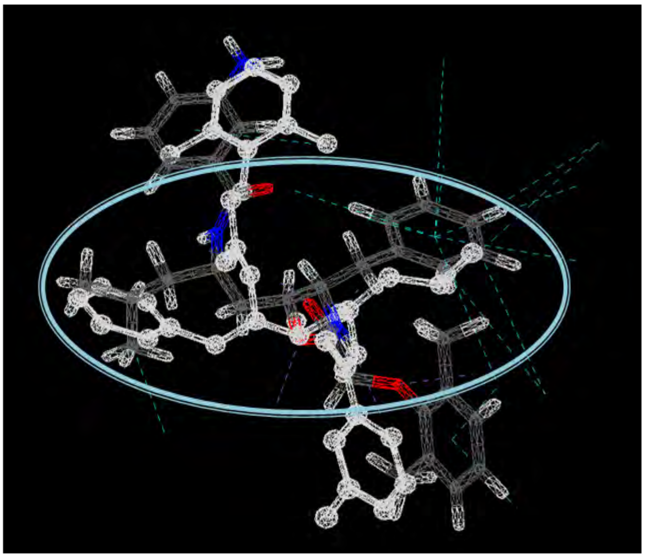
Figure 21: Re-docking of ligand (R36) into target structure 1LEE in parameter set 1 (top) and parameter set 2 (bottom).

The docking pose is represented in CPK color while the co-crystallized pose of R36 before docking is shown in white color. Interactions to the key residues are indicated in red circles on the text file (left hand side). Figure is generated by using FlexV.

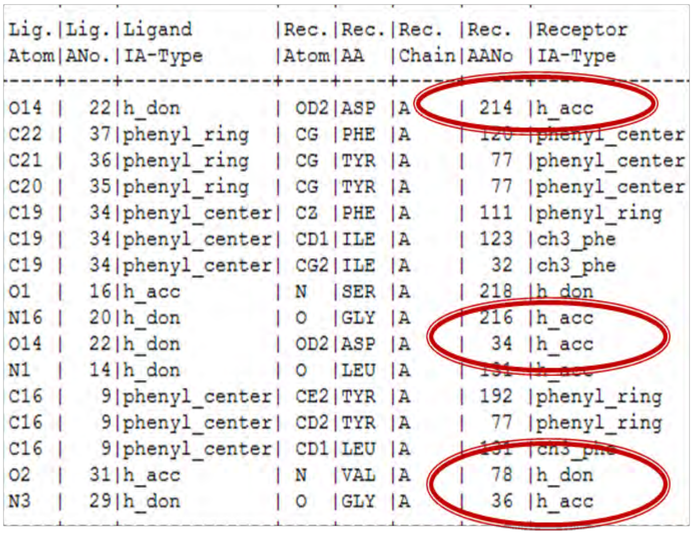
Parameter set 3



```
----+----+-------------+----+----+-----+-----+-------------
|Lig.|Lig.|Ligand       |Rec.|Rec.|Rec. |Rec. |Receptor
|Atom|ANo.|IA-Type      |Atom|AA  |Chain|AANo |IA-Type
----+----+-------------+----+----+-----+-----+-------------
|C21 |  36|phenyl_ring  | CG |PHE |A    | 294 |phenyl_center
|C21 |  36|phenyl_ring  | CG |TYR |A    | 192 |phenyl_center
|C19 |  34|phenyl_center| CD1|ILE |A    | 212 |ch3_phe
|C19 |  34|phenyl_center| CE2|TYR |A    | 192 |phenyl_ring
|C19 |  34|phenyl_center| CD2|LEU |A    | 292 |ch3_phe
|C7  |  28|ch3_phe      | CG |TYR |A    |  77 |phenyl_center
|C   |   1|phenyl_center| CE2|TYR |A    | 192 |phenyl_ring
|C   |   1|phenyl_center| CD1|LEU |A    | 131 |ch3_phe
|C16 |   9|phenyl_center| CG2|THR |A    | 217 |ch3_phe
|N3  |  29|h_don        | O  |GLY |A    | 216 |h_acc
|O14 |  22|h_don        | OD1|ASP |A    | 214 |h_acc
|O14 |  22|h_acc        | OD1|ASP |A    |  34 |h_don
|O1  |  16|h_acc        | OH |TYR |A    | 192 |h_don
|O26 |  19|h_acc        | N  |VAL |A    |  78 |h_don
|N16 |  20|h_don        | O  |GLY |A    |  36 |h_acc
|C22 |  37|phenyl_ring  | CG |PHE |A    | 294 |phenyl_center
|C19 |  34|phenyl_center| CG2|VAL |A    |  78 |ch3_phe
|C19 |  34|phenyl_center| CG1|VAL |A    |  78 |ch3_phe
|C19 |  34|phenyl_center| CD1|ILE |A    | 300 |ch3_phe
|C20 |  35|phenyl_ring  | CG |TYR |A    | 192 |phenyl_center
|C19 |  34|phenyl_center| CE1|PHE |A    | 294 |phenyl_ring
----+----+-------------+----+----+-----+-----+-------------
```



Parameter set 4



```
Lig.|Lig.|Ligand       |Rec.|Rec.|Rec. |Rec. |Receptor
Atom|ANo.|IA-Type      |Atom|AA  |Chain|AANo |IA-Type
----+----+-------------+----+----+-----+-----+-------------
O14 |  22|h_don        | OD2|ASP |A    | 214 |h_acc
C22 |  37|phenyl_ring  | CG |PHE |A    | 120 |phenyl_center
C21 |  36|phenyl_ring  | CG |TYR |A    |  77 |phenyl_center
C20 |  35|phenyl_ring  | CG |TYR |A    |  77 |phenyl_center
C19 |  34|phenyl_center| CZ |PHE |A    | 111 |phenyl_ring
C19 |  34|phenyl_center| CD1|ILE |A    | 123 |ch3_phe
C19 |  34|phenyl_center| CG2|ILE |A    |  32 |ch3_phe
O1  |  16|h_acc        | N  |SER |A    | 218 |h_don
N16 |  20|h_don        | O  |GLY |A    | 216 |h_acc
O14 |  22|h_don        | OD2|ASP |A    |  34 |h_acc
N1  |  14|h_don        | O  |LEU |A    | 131 |h_acc
C16 |   9|phenyl_center| CE2|TYR |A    | 192 |phenyl_ring
C16 |   9|phenyl_center| CD2|TYR |A    |  77 |phenyl_ring
C16 |   9|phenyl_center| CD1|LEU |A    | 131 |ch3_phe
O2  |  31|h_acc        | N  |VAL |A    |  78 |h_don
N3  |  29|h_don        | O  |GLY |A    |  36 |h_acc
```

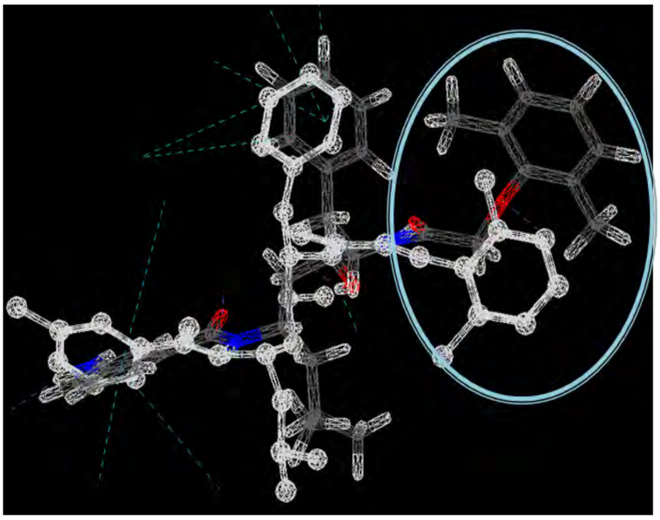


Figure 22: Re-docking of ligand (R36) into target structure 1LEE in parameter set 3 (top) and parameter set 4 (bottom).

The docking pose is represented in CPK colors while the co-crystallized pose of R36 before docking is shown in white color. Interactions to the key residues are indicated in red circles on the text file (left hand side). Deviations are highlighted in circles on the image (right hand side). Figure is generated by using FlexV.

	ASP 214	ASP 34	Val 78	Gly 36	Binding mode
Param 1	0	1	1	1	Good
Param 2	0	1	1	1	OK
Param 3	1	1	1	1	Twisted
Param 4	1	1	1	1	OK

Table 12: Displays interaction information with significant amino acids for 1LEE and its co-crystallized ligand (R36) under different parameter sets. Significant amino acids are displayed along with the comment on the binding mode observed.

Deployment strategy

Several test runs were launched before the deployment docking jobs on the EGEE Grid. The major purpose of performing test runs is to cover technical as well as modeling aspects: the system credibility in terms of modeling, performance, setting and tuning of parameters, analyzing software parameter influence on docking results are checked.

As the X-ray resolution of target structures 1LEE (1.8 Å) and 1LF2 (1.9 Å) were well suited for our purposes, major test runs were performed on 1LEE and 1LF2. Test runs were performed with a combination of known compounds, (Walter Reed compounds [215] found to have micro molar inhibitions *in vitro*), randomly chosen 20,000 ZINC compounds, 400 ZINC compounds and most importantly on the FlexX 200 standard benchmark dataset [214].

Final Deployment

Based on the results from direct docking, re-docking, cross-docking and test runs, different parameters were prepared for both targets and docking software. Finally, large-scale computations have been performed on 8 receptor scenarios for FlexX and 10 receptor scenarios for AutoDock. One million compounds were screened against each receptor scenario on the EGEE Grid infrastructure. The data challenge witnessed 42 million docking experiments on more than 1,700 computers distributed in Europe and Asia. The details about Grid implementation and deployment of the docking jobs are out of scope of this thesis and can be found in [145, 146].

4.5 Results and Discussion

The AutoDock docking tool was applied in parallel to FlexX for the screening, but the final results of AutoDock were not convincing. This is due to some errors that occurred during the parameterisation of the AutoDock software parameters. Most of compounds docked by using AutoDock have high internal free energies; this is because of the autotors tool. Autotors is a part of AutoDock program and is used for the preparation of the compounds, especially while assigning torsions at the rotatable bonds.

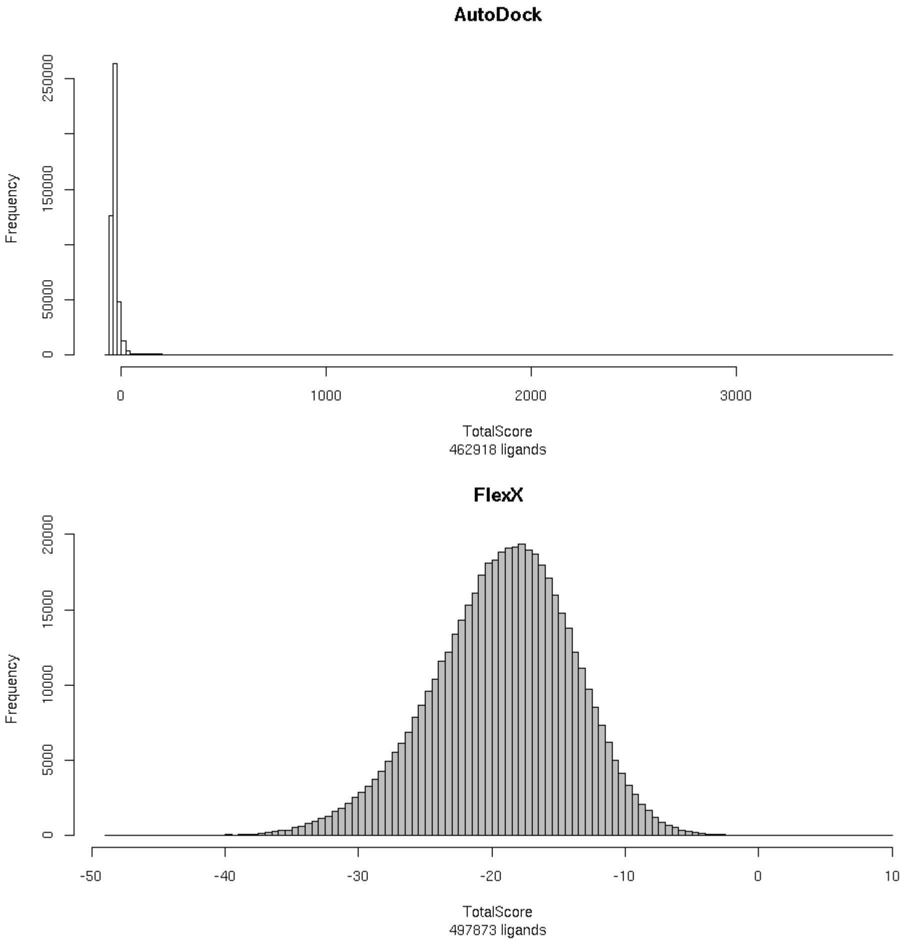


Figure 23: Score distribution plots of the AutoDock and FlexX in histogram representation. AutoDock scores (upper) and FlexX scores (lower).
It is clear from the plots that there is normal distribution of scores observed with FlexX software, whereas with AutoDock, the score distribution is quite abnormal and even some compounds scored in positive value.

Figure 23 displays a histogram plot, number of compounds (X-axis) against docking score (Y-axis, for the same parameter set. While, the FlexX scores are almost following a normal distribution, the AutoDock scores are distorted. The majority of ligands achieved a negative score, similarly distributed like the FlexX scores, but there are some ligands achieving unusually high scores with AutoDock.

Further analysis showed that these high energies stem from the internal stress term of the ligands, which AutoDock incorporates in its final docking score. The intermolecular energy was quite sensible. A deeper look at the structure of some of these ligands revealed that their pdbq files were not generated correctly by autotors. (Autotors is an executable of AutoDock suite, which converts mol2 file format to pdbq, script used for converting mol2 to pdbq is provided in Appendix). Since AutoDock uses the Van der Waals potential to model the internal ligand energies, this resulted in clashes and very high van der Waals values.

Summary of the output

As the results of AutoDock were not convincing, the subsequent analysis focused solely on the results obtained using FlexX. The outputs of the docking results in FlexX are log files. The log files are converted into CSV file format (comma-separated files). These CSV files in turn serve as input for VS explorer**,** a java based prototype software for analyzing virtual and high throughput screening results, developed at Fraunhofer-SCAI (www.scai.fraunhofer.de). Three different forms of results are saved and analyzed from each docking assay:

i. Docking scores of the ten best solutions after clustering.

ii. Interaction information between protein and ligands of the ten best solutions.

iii. Binding modes of the ten best solutions.

Moreover, ranking process is the integral part of the docking software. FlexX have a post processing optimization of the docking solution and clustering. Clustering in FlexX is based on RMSD, angle and distance deviation. Default values of FlexX are used as clustering cutoffs. The overall filtering process we employed is shown in Figure 24.

Clustering and match information

Usually, result analysis concentrates on the best ranking solution only or on the best 5 solutions based on docking score. However, when exhaustive analysis is done for all predictions there is a smoothening of score observed from the best solution to the next best solution in many cases. Moreover, the binding modes are very often nearly the same. To address this problem, result analysis of the ten best solutions after clustering is done: this allows screening diverse binding modes and identifying compounds with interactions to key residues of the protein even if the score is not optimal.

Strategies in result analysis

The aim of the result analysis is to reduce the number of false positives and finally identifying a few hundred promising compounds that can be tested in experimental laboratories. Results are analyzed at three levels (Figure 24 illustrates the overall filtering process):

1. By docking score
2. FlexX has a unique ability to compute the atom-to-atom interaction between the protein and the ligand. This information is exploited, and further used in analyzing results (match-information).
3. Manual visualization of binding modes inside the active site of the protein.

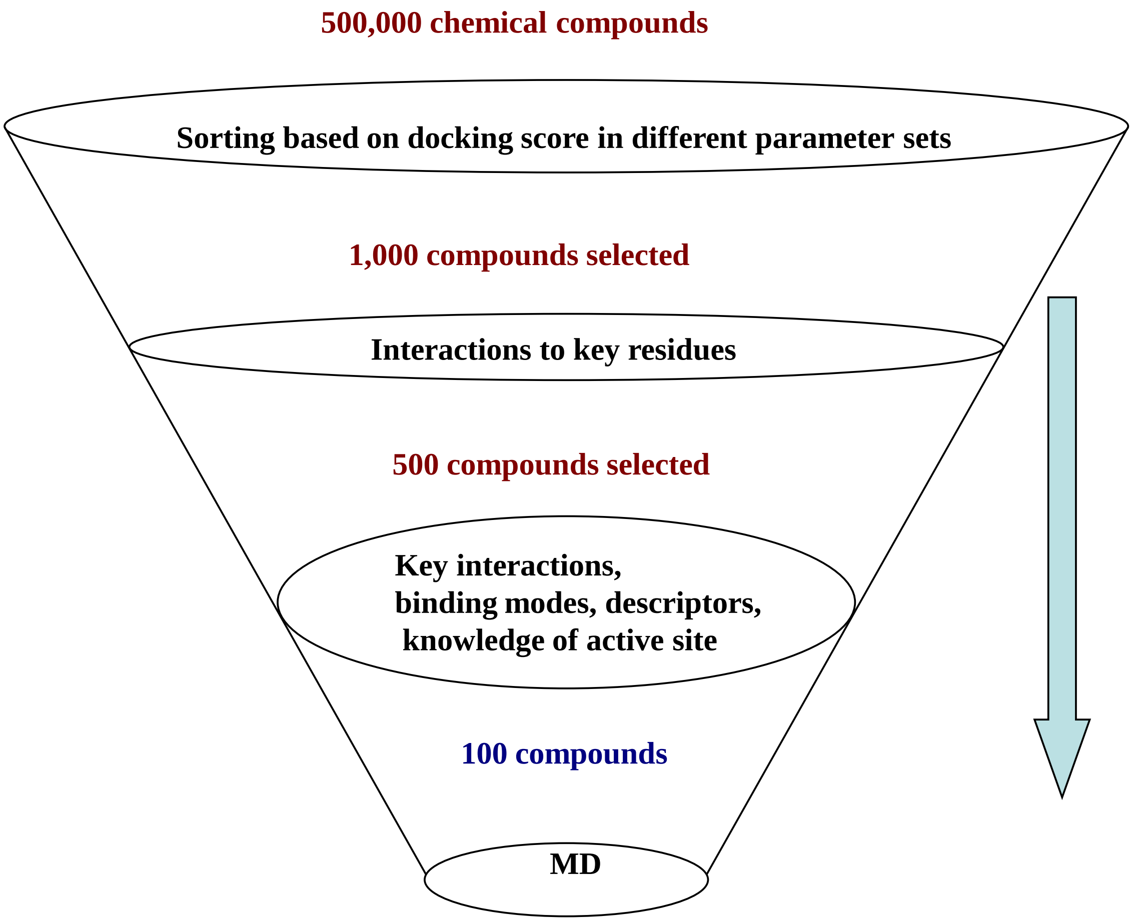

500,000 chemical compounds
Sorting based on docking score in different parameter sets
1,000 compounds selected
Interactions to key residues
500 compounds selected
Key interactions,
binding modes, descriptors,
knowledge of active site
100 compounds
MD

Figure 24: Representation of overall filtering process employed in WISDOM-I.
The process starts with 500,000 compounds and at different stages various filters applied as indicated in the figure to reduce the false positives and finally to identify the best possible hits which are more likely to be leads.

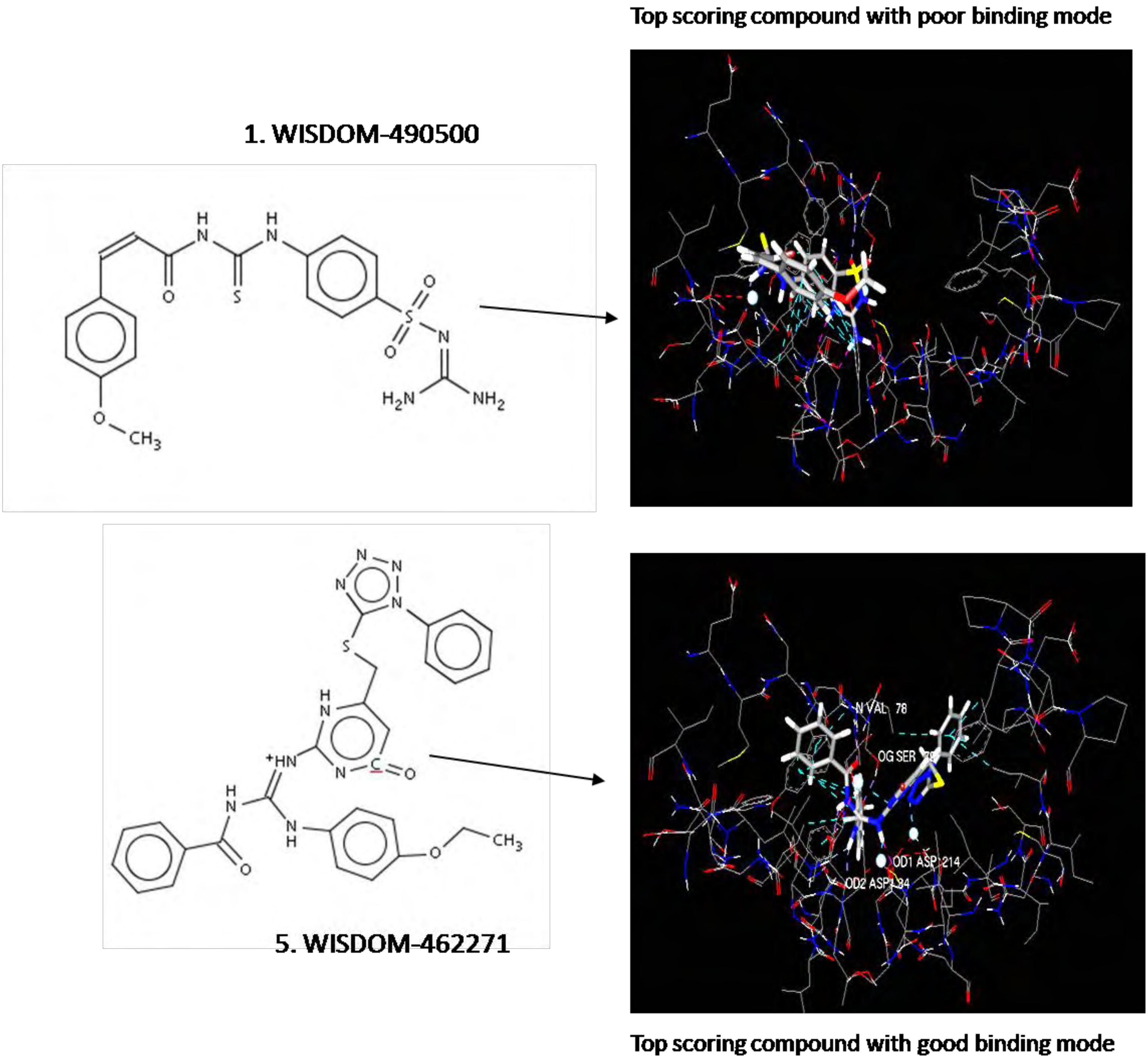
4.5.1 Top scoring compounds

As evident from the re-docking experiments, the target 1LEE without any inclusion of crystal water performed well both in terms of RMS deviations and docking scores, so we restricted the final result analysis to target structure 1LEE i.e., without any crystal water molecules. Figure 25 represents the top two compounds based on the score from parameter set 1. Visualization of the docking solutions shows that, although some compounds possess high scores they are quite far away from the center of the binding pocket. As we are looking for competitive inhibitors, the ideal compounds will be the ones, which are well within the binding pocket. Consequently, compounds, which are remote from the binding pocket, are rejected. An example of a top scoring compound with poor binding mode is represented in Figure 25.

Special attention has been given to all individual complexes for the top 1,000 compounds from all the four-parameter sets. Similar to the results obtained from parameter set 1, some of the top scoring compounds from various other parameter sets also failed to form the expected interactions to key residues of the protein, and further the binding mode of the docking poses inside the active site of 1LEE was not convincing. Therefore, we employed several filters for the final selection of the compounds (Figure 24 represents the filtering criteria employed). After undergoing the filtering procedure represented in Figure 24, finally, 100 chemical compounds have been selected for re-ranking by molecular dynamics.

The 100 compound list contains

- 20 compounds of Guanidino scaffold
- 20 compounds of Thiourea core scaffold
- 20 compounds of Urea core scaffold
- 30 compounds with diverse scaffold
- 10 low scoring compounds (for control studies)



Top scoring compound with poor binding mode

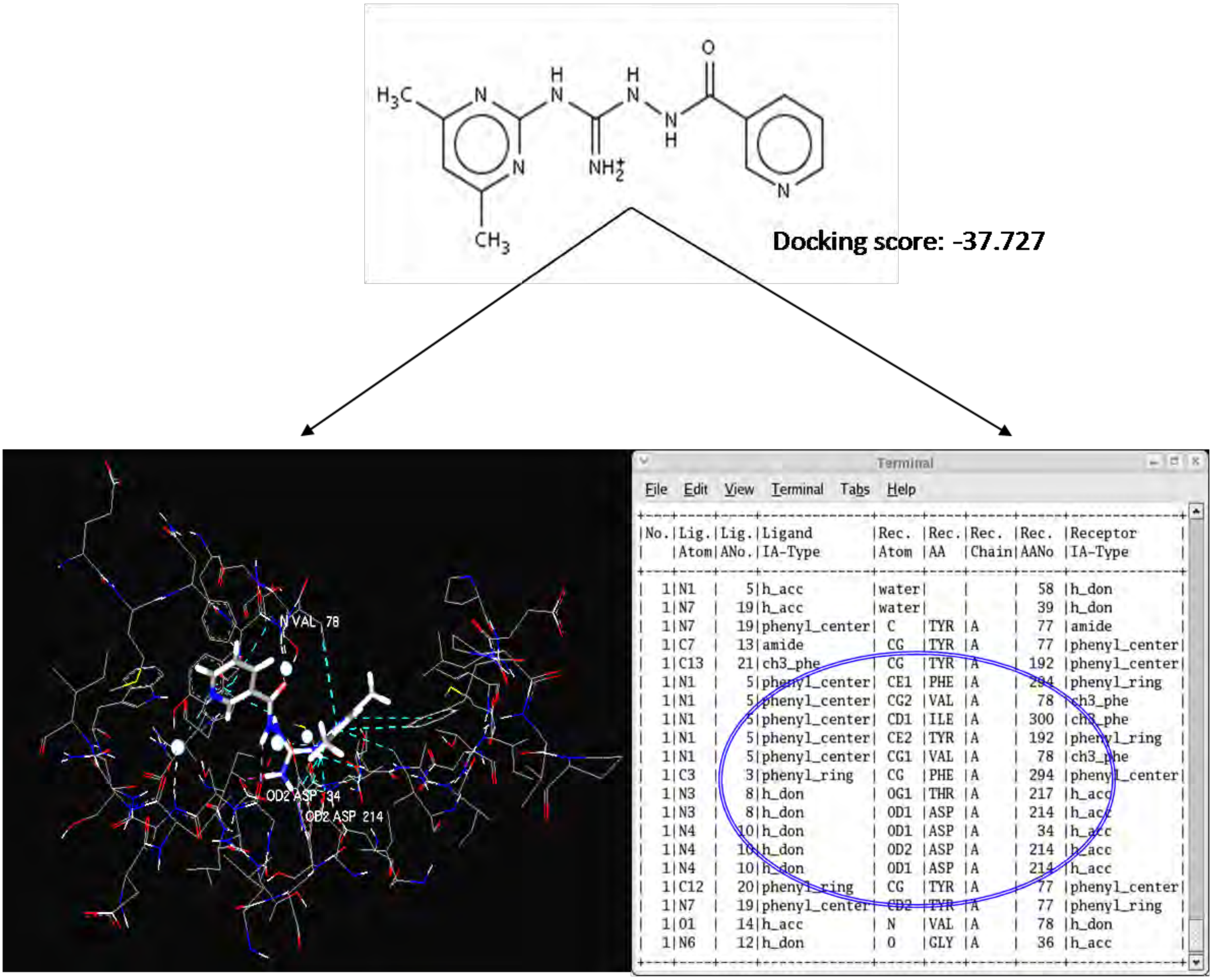
1. WISDOM-490500

5. WISDOM-462271

Top scoring compound with good binding mode

Figure 25: Representation of the top scoring compounds in parameter set 1.
Top scoring compounds with poor binding mode and good binding modes inside the active site of the protein (pdb id: 1LEE) are indicated with arrows.

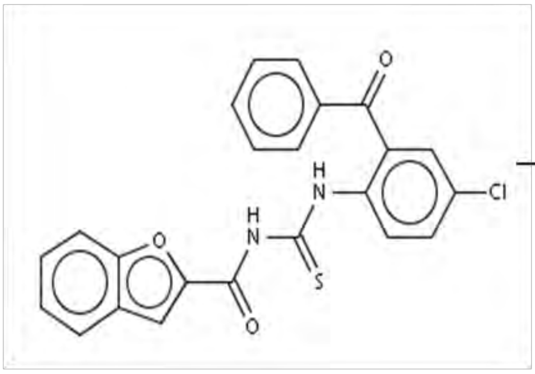
Guanidino analogue

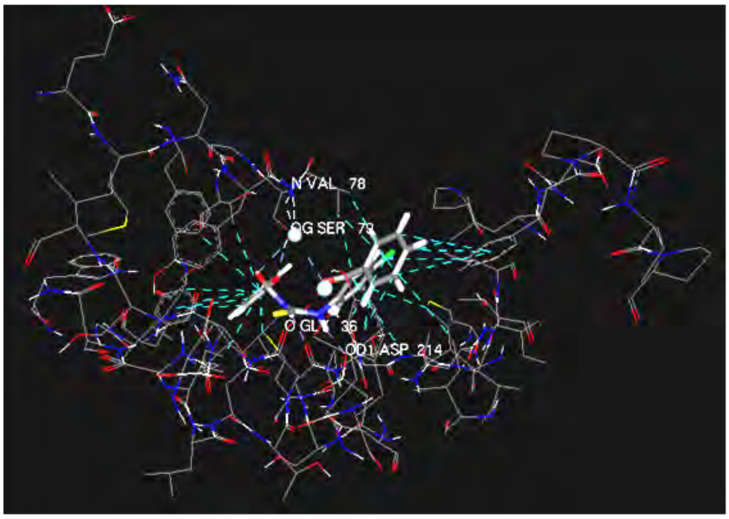


```
+---+----+----+-------------+-----+----+-----+-----+-------------+
|No.|Lig.|Lig.|Ligand       |Rec. |Rec.|Rec. |Rec. |Receptor     |
|   |Atom|ANo.|IA-Type      |Atom |AA  |Chain|AANo |IA-Type      |
+---+----+----+-------------+-----+----+-----+-----+-------------+
|  1|N1  |   5|h_acc        |water|    |     |  58 |h_don        |
|  1|N7  |  19|h_acc        |water|    |     |  39 |h_don        |
|  1|N7  |  19|phenyl_center| C   |TYR |A    |  77 |amide        |
|  1|C7  |  13|amide        | CG  |TYR |A    |  77 |phenyl_center|
|  1|C13 |  21|ch3_phe      | CG  |TYR |A    | 192 |phenyl_center|
|  1|N1  |   5|phenyl_center| CE1 |PHE |A    | 294 |phenyl_ring  |
|  1|N1  |   5|phenyl_center| CG2 |VAL |A    |  78 |ch3_phe      |
|  1|N1  |   5|phenyl_center| CD1 |ILE |A    | 300 |ch3_phe      |
|  1|N1  |   5|phenyl_center| CE2 |TYR |A    | 192 |phenyl_ring  |
|  1|N1  |   5|phenyl_center| CG1 |VAL |A    |  78 |ch3_phe      |
|  1|C3  |   3|phenyl_ring  | CG  |PHE |A    | 294 |phenyl_center|
|  1|N3  |   8|h_don        | OG1 |THR |A    | 217 |h_acc        |
|  1|N3  |   8|h_don        | OD1 |ASP |A    | 214 |h_acc        |
|  1|N4  |  10|h_don        | OD1 |ASP |A    |  34 |h_acc        |
|  1|N4  |  10|h_don        | OD2 |ASP |A    | 214 |h_acc        |
|  1|N4  |  10|h_don        | OD1 |ASP |A    | 214 |h_acc        |
|  1|C12 |  20|phenyl_ring  | CG  |TYR |A    |  77 |phenyl_center|
|  1|N7  |  19|phenyl_center| CD2 |TYR |A    |  77 |phenyl_ring  |
|  1|O1  |  14|h_acc        | N   |VAL |A    |  78 |h_don        |
|  1|N6  |  12|h_don        | O   |GLY |A    |  36 |h_acc        |
+---+----+----+-------------+-----+----+-----+-----+-------------+
```

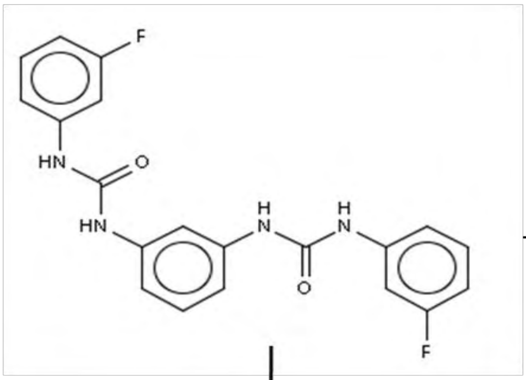
Figure 26: Representation of one of the top scoring guanidino analogue.
The docking score (kJ/Mol) with FlexX is indicated below the compound. The compound exibits ideal binding mode and interactions to key residues. The interactions between the compounds and the key amino acids of the protein are highlighted in blue circle.

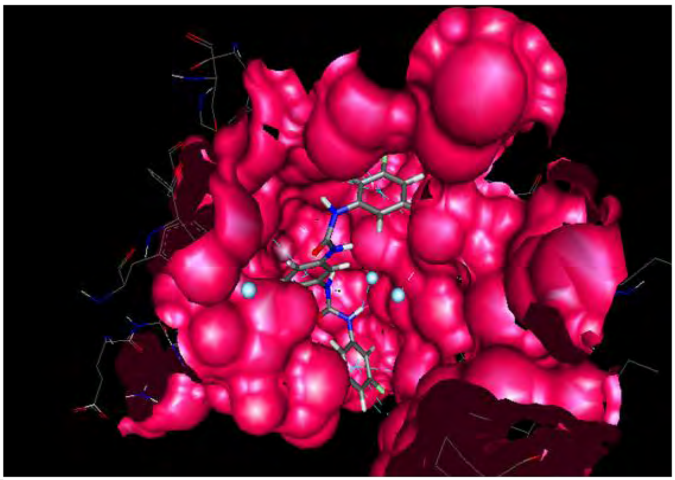
A. Thiourea anlogue

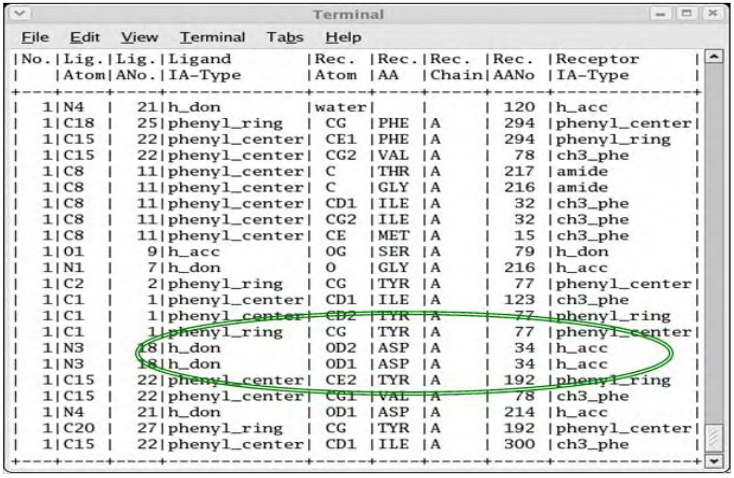




B. diphenyl urea analogue







```
|No.|Lig.|Lig.|Ligand          |Rec. |Rec.|Rec. |Rec. |Receptor      |
|   |Atom|ANo.|IA-Type         |Atom |AA  |Chain|AANo |IA-Type       |
+---+----+----+----------------+-----+----+-----+-----+--------------+
|  1|N4  |  21|h_don           |water|    |     | 120 |h_acc         |
|  1|C18 |  25|phenyl_ring     | CG  |PHE |A    | 294 |phenyl_center|
|  1|C15 |  22|phenyl_center|  CE1 |PHE |A    | 294 |phenyl_ring   |
|  1|C15 |  22|phenyl_center|  CG2 |VAL |A    |  78 |ch3_phe       |
|  1|C8  |  11|phenyl_center|  C   |THR |A    | 217 |amide         |
|  1|C8  |  11|phenyl_center|  C   |GLY |A    | 216 |amide         |
|  1|C8  |  11|phenyl_center|  CD1 |ILE |A    |  32 |ch3_phe       |
|  1|C8  |  11|phenyl_center|  CG2 |ILE |A    |  32 |ch3_phe       |
|  1|C8  |  11|phenyl_center|  CE  |MET |A    |  15 |ch3_phe       |
|  1|O1  |   9|h_acc           | OG  |SER |A    |  79 |h_don         |
|  1|N1  |   7|h_don           | O   |GLY |A    | 216 |h_acc         |
|  1|C2  |   2|phenyl_ring     | CG  |TYR |A    |  77 |phenyl_center|
|  1|C1  |   1|phenyl_center|  CD1 |ILE |A    | 123 |ch3_phe       |
|  1|C1  |   1|phenyl_center|  CD2 |TYR |A    |  77 |phenyl_ring   |
|  1|C1  |   1|phenyl_ring     | CG  |TYR |A    |  77 |phenyl_center|
|  1|N3  |  18|h_don           | OD2 |ASP |A    |  34 |h_acc         |
|  1|N3  |  18|h_don           | OD1 |ASP |A    |  34 |h_acc         |
|  1|C15 |  22|phenyl_center|  CE2 |TYR |A    | 192 |phenyl_ring   |
|  1|C15 |  22|phenyl_center|  CG1 |VAL |A    |  78 |ch3_phe       |
|  1|N4  |  21|h_don           | OD1 |ASP |A    | 214 |h_acc         |
|  1|C20 |  27|phenyl_ring     | CG  |TYR |A    | 192 |phenyl_center|
|  1|C15 |  22|phenyl_center|  CD1 |ILE |A    | 300 |ch3_phe       |
+---+----+----+----------------+-----+----+-----+-----+--------------+
```

Figure 27: (A) Top scoring thiourea analogue. (B) Top scoring diphenyl urea analogue.
The binding modes inside the activesite and interaction information are shown. Highlighted in green circle are compounds interactions to one of the catalytic residue (ASP 34) of the target protein. Figure generated by using FlexV. Interactions are indicated in dotted lines

Repeated identification of diphenyl urea analogues in the top 1000 compounds by docking score, which incidentally happens to be known inhibitors against plamsepsin [215] (Walter Reed compounds) suggests that the approach is valid and sensible. Figure 27 (B) represents a diphenyl urea analogue inside the active site and the interactions to key residues of the target are highlighted in a circle. A close inspection revealed that this compound displays a similar binding mode as co-crystallized ligand (R36) and forms the required interactions to key residues of plasmepsin.

The other group of compounds gathers thiourea analogues (Figure 27(A)). There is always a consensus in placing the core group (thiourea) with the sulphur atom positioning itself towards the flap residue Val78, and the two nitrogen atoms making interactions to the catalytic Asp214 or Asp34 or both. This observation is well in concordance with the binding modes of the Walter Reed compounds and binding modes of the co-crystallized ligands.

The most significant observation from the current study is the identification of the guanidino analogues. These compounds are very promising as they obeyed all the filtering criteria employed in finding the hits. Figure 26 represents the guanidino analogues with their respective docking scores in kJ/Mol. The binding mode of a guanidino derivative is shown in the active site of protein 1LEE. The interactions to key residues are highlighted in circles. Similar to the binding modes of thiourea compounds, there was a consensus observed in the binding modes of guanidino compounds: the deprotonated nitrogen atom positioning itself towards the flap residue Val78 and the adjacent nitrogen atoms making interactions to catalytic residues Asp214 and/or Asp34 (Figure 26).

Guanidino analogues are likely to be a novel class of compounds, as they have not yet reported as inhibitors for plasmepsins. Additionally, chemically diverse compounds have also been identified as hits, including thiazole analogues.

Eighteen different chemical descriptors are calculated for all the finally selected 100 compounds to reduce the late stage attrition rates. However, all the compounds possess acceptable chemical descriptor values. Figure 28 displays the top hundred compounds and their chemical descriptors. Table 13 displays the docking score and chemical descriptor values of all the hundred compounds.

Results of top 100 compounds and their descriptor values

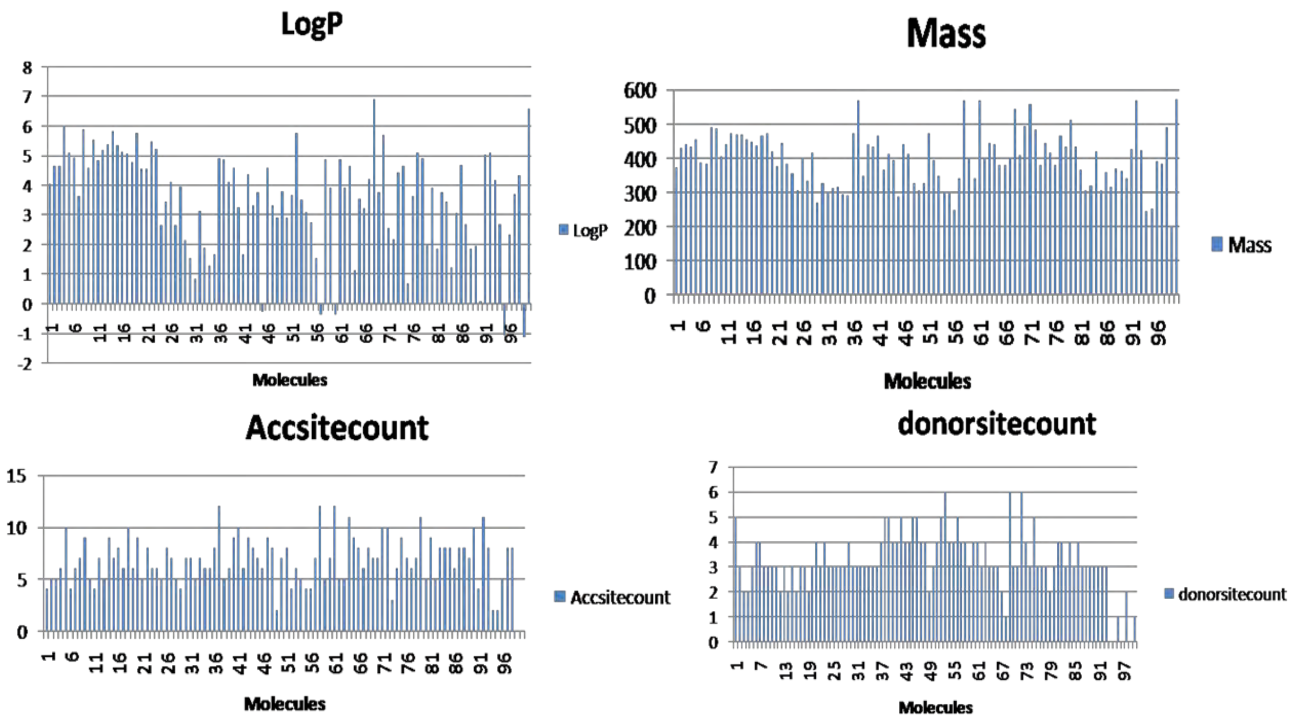


Figure 28: Top hundred compounds and their chemical descriptor values.
Very important chemical descriptors such as LogP, molecular weight (top), H-bond donor, H-bond acceptor (bottom) are plotted. These plots demonstrate that all the compounds possess acceptable chemical descriptor values.

Table 13: Displays best 100 compounds that were selected against plasmepsin from the large-scale virtual screening of 500,000 compounds.
This Table displays docking scores and important chemical descriptor values for these compounds

Molecule	WISDOM					
No.	ID	Score	Mass	H-Acc	H-Don	LogP
1	280991	-38.763	373.305	4	5	4.05
2	380406	-38.103	429.449	5	3	4.65
3	378548	-39.747	441.506	5	2	4.65
4	193748	-38.285	434.896	6	2	6
5	242452	-40.407	454.887	10	3	5.09
6	313614	-38.034	388.46	4	4	4.95
7	312057	-38.412	382.41	6	4	3.64
8	384677	-39.681	489.974	7	3	5.89
9	310954	-38.2	487.756	9	3	4.58
10	243118	-37.174	403.498	5	3	5.55
11	382373	-37.633	439.51	4	3	4.86
12	385534	-37.723	471.507	7	2	5.21
13	372757	-37.082	468.324	5	3	5.39
14	373697	-40.574	466.898	9	2	5.83

15	373762	-37.302	453.942	7	3	5.35
16	242449	-38.131	446.48	8	2	5.12
17	492970	-39.791	435.54	6	3	5.06
18	475515	-39.502	464.495	10	3	4.79
19	404128	-37.418	470.544	6	2	5.76
20	326015	-38.106	420.442	9	3	4.57
21	329771	-37.112	375.445	5	4	4.55
22	386759	-39.937	443.479	8	3	5.5
23	430276	-37.045	382.364	6	4	5.24
24	313546	-36.875	353.329	6	3	2.65
25	109865	-37.293	303.743	5	3	3.44
26	416361	-36.661	398.239	8	3	4.12
27	120595	-36.37	333.726	7	3	2.65
28	437779	-36.029	415.613	5	3	3.96
29	89351	-36.869	269.299	4	4	2.13
30	178145	-36.123	325.15	7	3	1.52
31	170421	-35.821	300.27	7	3	0.82
32	178319	-36.553	313.351	5	3	3.13
33	170305	-36.106	315.299	7	3	1.9
34	73901	-40.888	293.318	6	3	1.29
35	81354	-37.093	290.705	6	3	1.66
36	315095	-43.098	472.923	8	3	4.92
37	462971	-45.625	568.631	12	4	4.87
38	52923	-44.035	349.41	5	5	4.11
39	261841	-43.119	440.477	6	5	4.6
40	300822	-41.736	434.468	9	4	3.24
41	305608	-41.596	464.451	10	4	1.65
42	392786	-40.486	365.194	6	5	4.38
43	316830	-38.111	412.849	9	4	3.32
44	396606	-37.822	392.455	8	4	3.76
45	17970	-37.727	286.313	7	5	-0.29
46	261841	-43.119	440.477	6	5	4.6
47	491148	-38.111	412.849	9	4	3.32
48	49805	-39.53	326.287	8	4	2.89
49	30030	-18.408	306.385	2	2	3.8
50	495606	-35.979	328.389	7	3	2.91
51	306328	-35.293	470.5	8	4	3.68
52	301748	-35.33	393.53	4	5	5.78
53	141813	-35.918	349.41	6	6	3.5
54	135119	-35.781	299.28	5	4	3.08
55	497564	-35.585	297.289	4	4	2.73
56	128986	-34.506	248.261	4	5	1.53
57	175272	-45.827	341.361	7	4	-0.38
58	462971	-45.625	568.631	12	4	4.87

59	281783	-36.953	398.437	5	3	3.91
60	175272	-45.827	341.361	7	4	-0.38
61	462971	-45.625	568.631	12	4	4.87
62	281783	-36.953	398.437	5	3	3.91
63	313358	-42.483	444.53	5	4	4.66
64	475829	-39.296	439.489	11	3	1.11
65	426087	-38.169	378.787	9	3	3.55
66	412044	-37.943	380.374	8	3	3.22
67	208216	-37.773	400.405	6	2	4.2
68	479589	-37.665	543.678	8	1	6.93
69	316673	-41.031	409.483	7	6	3.78
70	378421	-43.771	493.553	7	3	5.7
71	229724	-36.073	555.624	10	3	2.54
72	252811	-38.984	484.507	10	6	2.18
73	358887	-36.85	379.437	3	4	4.44
74	313357	-41.931	443.522	6	3	4.66
75	497987	-40.011	414.435	9	5	0.66
76	262376	-30.361	380.251	7	3	3.64
77	260587	-33.101	463.511	6	3	5.11
78	300002	-31.182	434.342	7	3	4.9
79	259529	-34.998	511.529	11	2	1.98
80	253632	-30.398	434.489	5	3	3.93
81	253622	-35.441	365.343	9	4	1.85
82	107022	-33.179	306.339	5	4	3.78
83	92393	-34.368	320.708	8	3	3.45
84	223835	-32.364	417.288	8	4	1.23
85	92712	-32.682	304.253	8	3	3.07
86	202102	-33.614	358.802	6	4	4.69
87	141291	-31.029	316.289	8	3	2.68
88	290905	-34.233	369.399	8	3	1.84
89	295687	-32.286	361.369	7	3	1.96
90	67405	-34.471	342.313	10	3	0.07
91	193711	-36.11	425.484	4	3	5.04
92	245953	-35.249	566.395	11	3	5.11
93	295689	-29.729	422.478	8	3	4.19
94	66304	-8.628	243.344	2	0	2.68
95	89585	-6.027	250.405	2	0	-1.09
96	420105	-10.749	390.271	5	1	2.34
97	345897	-12.398	383.462	8	0	3.69
98	207090	-14.147	490.978	8	2	4.34
99	437455	1.168	198.368	0	0	-1.14
100	456497	-2.665	570.758	0	1	6.61

4.6 Summary

This chapter describes the complete setup, and large-scale virtual screening effort for finding novel compounds active against *Plasmodium falciparum* plasmepsins. The screening of compounds was performed on the EGEE Grid infrastructure (Grid details given elsewhere [146, 145]). The screening effort presented in this chapter is a part of WISDOM project, and is one example for the successful utilization of the e-Science paradigm in the area of computational life sciences. Making use of the world's largest scientific compute infrastructure, the EGEE Grid, we have realized a large virtual screening project aiming at the identification of new, potential candidate molecules against the plasmepsin family of aspartic proteases encoded by *Plasmodium falciparum,* the malaria causing protozoan parasite. Besides the demonstration that a global e-Science production infrastructure such as EGEE, with the identification of new family of potential inhibitors, the guanidino group of compounds, we have established a new class of chemical entities with inhibitory activity against *Plasmodium falciparum* plasmepsins. A strong support for their putative activity is that most of the so far known antimalarial drugs likewise contain basic groups. The virtual screening approach taken by us could be subject to criticism as alternative strategies such as pharmacophore or similarity searches do exist. However, the fact that we were able to point to a new class of potential inhibitors after using a selection of publicly available "virtual" compounds (the ZINC database of compounds) and the fact that we could identify candidate inhibitors that fall into the already well-established inhibitor classes of thiourea and diphenyl urea analogues speak for the route we have taken.

Several potential issues were identified during the large-scale docking experiment both on the technical side and on modeling side. On the technical side, handling massive docking data as flat files was a huge challenge. This is one of the reasons, why result analyses were not performed on all the variations of protein and parameter sets. A customized docking database (such as docking database (DDB) from BioSolveIT Gmbh) would be an ideal choice for storing and analyzing the results.

On the modeling side, due to the robust nature of the docking algorithm and scoring function, significant parameters such as protein flexibility and solvent parameters were ignored. This may be one of the reasons, why several compounds were having huge van der Waals clashes with the receptor atoms. Other issues include the orientation of peptide bond in docking conformations of thiourea and urea compounds. Peptide bond usually adopts planar orientations (180°) but in some cases, FlexX assigned 90°, which completely changed the hydrogen-bonding pattern of the compounds. To overcome the modeling issues, specifically the protein flexibility, solvent parameters and orientation deficiencies; molecular dynamics simulations were performed and is discussed in detail in chapter 5.

5 Chapter 5. Discovery of novel plasmepsin inhibitors by refining and rescoring through molecular dynamics

This chapter reports the significance of performing molecular dynamics and rescoring against the best hits resulting from docking experiment. The 5000 best-scoring conformations selected from the plasmepsin virtual screening described in chapter 4 were subjected to molecular dynamics simulations. The Amber software was used for molecular dynamics simulations; further rescoring was done applying MM-PBSA and MM-GBSA methods. Finally, this chapter reports the identification of novel small molecules and experimental results of the 30 novel compounds identified against plasmepsin II.[3]

5.1 Introduction

The significance of molecular dynamics simulations and rescoring the docking conformations by using sophisticated scoring functions are already discussed in chapter 2. Indeed compared to docking, molecular dynamics address the electrostatic solvation parameters, protein flexibility and additional degree of freedom. Consequently, it requires much higher time and computing power than docking, hence molecular dynamics can only be applied to a restricted number of compounds, usually the best hits coming out of the docking step. Therefore, molecular dynamics simulations appear to be very promising in improving the structure-based virtual screening process by addressing issues that were ignored by docking methods.

Molecular Dynamics (MD) analysis significantly changes the scoring of the best compounds, and it is therefore very important to apply it to a significant fraction of compounds, i.e., as many compounds as possible. Therefore, computational Grids appear very promising to improve the virtual screening process by increasing the number of compounds that will be processed by using molecular dynamics. The molecular dynamics deployment on the Grid is already described in chapter 3.

In context to the current thesis, extension of the docking application i.e., further re-scoring of plasmepsin docking conformations is needed, thus a refinement of the docking poses by molecular dynamics simulations has been implemented; the application of this extension of the workflow is described in the present chapter.

This chapter is based on G. Degliesposti, Kasam. V et al .Design and Discovery of Plasmepsin II Inhibitors Using an Automated Workflow on Large-Scale Grids. ChemMedChem, 2009, july;4(7):1164-73.

In molecular dynamics simulations, the automated relaying of output to the particular next stage and the management of the data emerging during the progress of calculations is complex, and is the main challenge. This is due to the nature of the software and the following procedure of file conversion, energy minimization, molecular dynamics simulations, and analysis of results. This dynamic behavior together with the sequential nature of interdependent jobs to be performed in a stepwise parallel fashion is the major problem in such a workflow.

To overcome these problems, handling several steps in sequential order without losing any significant information, Prof. Giulio Rastelli has proposed an automated workflow using the Amber software [216]. This procedure is suitable and highly configured for virtual screening purposes. Therefore, it is adapted and utilized to optimize the docking conformations that were resulted from the high throughput virtual screening experiments (described in chapter 4).

5.2 Rescoring by Amber software

Amber9

The latest version of the Amber software, Amber9 is used. The Amber9 is the suite of different tools which collectively carry out molecular dynamic simulations. The simulations by the Amber program can be divided into three phases, and each of these phases is performed by different executables of the Amber suite. Encoding these steps in separate programs has some important advantages. Firstly, it allows individual pieces to be up graded or replaced with minimal impact on other parts of the program suite.

Secondly, it allows different programs to be written with different coding practices. For example, LEAP is written in C using X-window libraries, ptraj and antechamber are text-based C codes, mm-pbsa is implemented in Perl, and the main simulation programs are coded in Fortran.

Thirdly, this separation often eases porting to new computing platforms: only the principal simulation codes (sander and pmemd) need to be coded for parallel operation or need to know about optimized (perhaps vendor-supplied) libraries. The general procedure employed by the Amber9 software is demonstrated in Figure 29.

A. Preparatory phase
B. Stimulatory phase
C. Analysis phase
D. Rescoring

A. Preparatory Phase

Antechamber

Antechamber is designed to be used with "General Amber Force Field" (GAFF). This force field is specifically designed to cover most of the parameters for chemical compounds and is compatible with traditional Amber force fields in such a way that both can be mixed during a simulation. Antechamber is used to assign atom types to molecules, and calculate a set of point charges. It can perform many file conversions and can also assign atomic charges and atom types. The output of antechamber is a "prepi file" which later serves as input for Leap program.

Antechamber solve the following problems:

1) Automatically identify bond and atom types

2) Judge atomic equivalence

3) Generate residue topology files

4) Find missing force field parameters and supply reasonable suggestions

Leap

Leap summarizes several small steps necessary to convert either the original input (protein.pdb and ligand.mol2) or output from a previous step (protein.pdb and ligand.mol2) into the sander-compatible input files: system.top and system.crd.

Since proteins are composed modularly of 20 different amino acids, tleap assigns atom types and partial charges automatically according to templates. Thus, protein.pdb-files can be read in immediately. For other organic molecules (ligands, cofactors), which are not included in the library of known templates, from the mol2-files the necessary atom types and partial charges have to be derived previously, thus, a hidden sequence of tools applies. Within the latter, essentially, antechamber produces a ligand.prep file. For assigning partial charges you can choose between two different charge methods:

a. Gasteiger charges: fast method.
b. AM1/BCC charges: slow, but robust method. Computes conformation dependent partial charges by semiempirical calculations using the program divcon).

In the current thesis, AM1/BCC charges are utilized.

B. Simulatory Phase

Sander

The acronym stands for Simulated Annealing with NMR-Derived Energy Restraints. The main program of Amber is sander. It carries out energy minimization, molecular dynamics and NMR refinements. It step performs force field calculations which are apt to modify the coordinates of the system (previously assembled by the tleap-step or a previous sander step). The immediate output of the Sander-step is the restart-file; its format however is not compatible to common viewing programs or e.g. FlexX or even Tleap. Thus, it has to be transformed into the handier output format (pdb) by another program called ambpdb.

C. Analysis Phase

Ptraj

Ptraj is a program to process and analyze sets of 3-D coordinates read in from a series of input coordinate files. It reads prmtop format in the Amber software. For each coordinate set read in, a sequence of events is performed (in a order specified) on each of the configuration (set of coordinates) read in. After processing all the configurations, a trajectory file and other supplementary data can be optionally written out. To use this program it is necessary to have

a. Parameter/topology file

b. List of input coordinate files

c. Optionally specify an output file

d. To specify a series of actions to be performed on each coordinate set read in

D. Rescoring by MM-PBSA and MM-GBSA

MM-PBSA calculations are performed with the executables integrated in the Amber9 software. It estimates energies and entropies from the snapshots contained within trajectory files (Created during the molecular dynamic simulations by Sander program). The calculations are organized and spawned by a Perl script; it collects statistics and formats the output in tabular form. The analysis is primarily based on continuum solvation models.

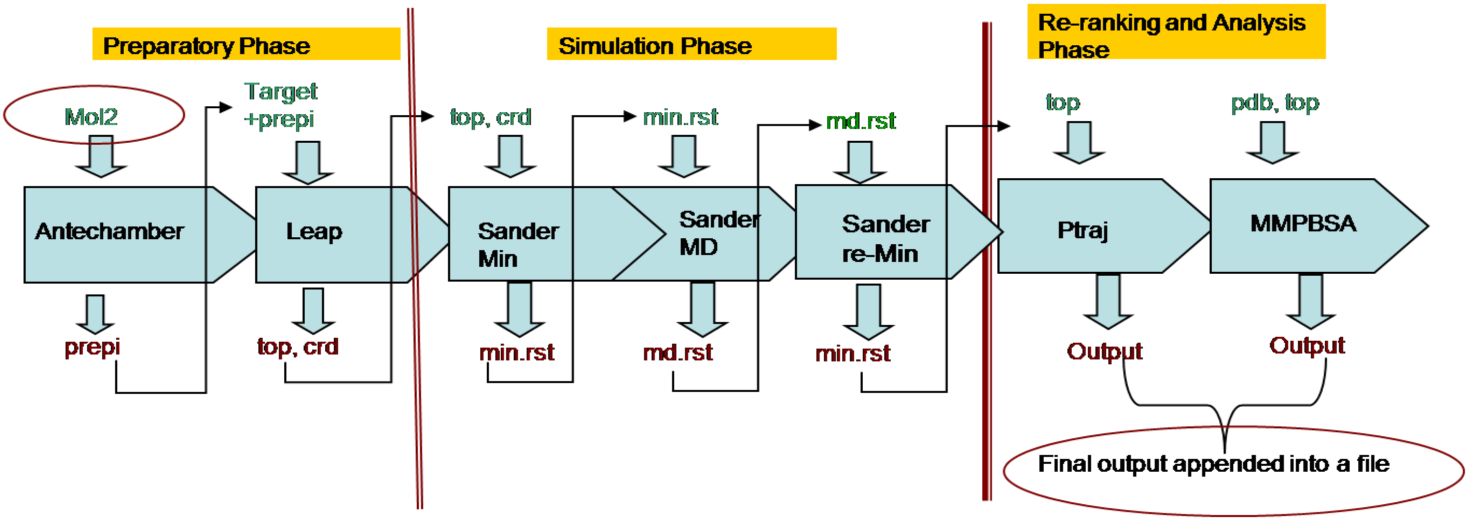


Figure 29 : General workflow of an Amber application.
Different phases of simulation and their associated toolsa nd their respective output file formats, are displayed. The Figure demonstates that output at one tool serves as an input for the next tool.

Theoretical back ground of MM-PBSA and MM-GBSA

The MM/PBSA method [217, 218] was introduced by Srinivasan et al [219]. This method proposes a post processing method to evaluate binding free energies in solution. It combines molecular mechanics energies with continuum solvent approaches to estimate binding free energies. The structures are usually collected from the molecular dynamics simulations or Monte Carlo methods. More detailed theoretical aspects of MM-PBSA and MM-GBSA can be found in [187, 188, 189]. The total binding free energy of the system can be calculated with the equation:

$$\Delta\, G_{\mathbf{binding}} = \Delta\, E_{\mathbf{MM}} + \Delta\, G_{\mathbf{solvent}} \;\; -T\,\Delta\, S_{\mathrm{solute}}$$

$$\Delta E_{MM} = \text{interaction energy} + \text{van der Waals} + \text{Electrostatics}$$

$$\Delta G_{\text{solvent}} = \Delta G_{\text{solv polar}} + \Delta G_{\text{solv non-polar}}$$

$$\text{PB} \qquad \text{GB} \qquad \text{SASA}_{\text{LCPO}}$$

Where, ΔE_{MM} is the molecular mechanics contribution expressed as the sum of the internal, electrostatics and van der Waals contributions to binding in dielectric constant,

ΔG_{solv} is the solvation free energy contribution to binding expressed as the sum of polar and nonpolar solvation free energies (ΔGsolv = ΔGpolar solv + ΔGnon-polar solv respectively);

ΔGsolvation, the solvation free energy, is calculated in two parts, the electrostatic component Gpolar using a Poisson–Boltzmann approach, and a non-polar part using the solvent-accessible surface area (SASA) model.

TΔSsolute is the contribution of solute entropy to binding and can be calculated by normal mode analysis [220].

The entropy change can be omitted, if only the relative binding energies of a series of structurally similar compounds are required, but if the absolute energy is important, or if the compounds are notably different, then its contribution to the final free energy cannot be ignored. A study by Kuhn et al [96], suggests that the MM-PBSA function could be used as a post-docking filter during the virtual screening of compounds, as their use of a single relaxed structure provided better results than usual averaging over MD simulation snapshots.

Ligand database and preparation

The prerequisite for refining and rescoring of compounds by molecular dynamics is; the compound, against which the simulation is performed, should be well inside the binding site of the receptor. For this reason, the docking conformations of best 5000 compounds against plamepsin (chapter 4) were taken as input. The docking scores of plasmepsin compounds when plotted against the number of compounds revealed that there is a significant rise in the score of the first 5000 compounds and from there on the docking scores were stable. Hence, the top scoring 5000 docking conformations were selected for molecular dynamics simulations.

Preparation of the mol2 database

AM1-BCC atomic charges were calculated for each compound, and sybyl atom types were assigned. A new mol2 file was written, and used for subsequent refinement. The protonation state of ligands was left unchanged. The Antechamber program of the Amber software performed the preparation of ligand database.

Preparation of plasmepsin II structure

Starting from the crystal structure of plasmepsin (pdb id: 1LEE), the structure was prepared for molecular dynamics simulation that is compatible with the Amber software. All the protonation states of the amino acids were assigned to a state consistent with pH 5, because plasmepsins are expressed in the acidic food vacuole (acidic pH conditions prevail). Aspartic and Glutamic acids were treated as deprotonated except otherwise noted. Lysine and Arginine residues were protonated. Histidines were treated as neutral, except for His164, tautomeric forms were assigned based on favourable hydrogen bonding with nearby residues. Evidence in the literature [221, 207, 222] point out that a few residues should be given special attention;

based on the literature, Asp34 and Asp303 were considered as neutral (COOH), and His164 was considered protonated. Asp34 is of special relevance, as it is one of the two aspartic acids in the catalytic dyad. Likewise, HIV protease, which shares many similarities with plasmepsins, are known to have one of the two Asp residues (sometimes both) protonated when inhibitors bind. It is anticipated that having a neutral Asp34 residue instead of a carboxylate would change the hydrogen bonding pattern of ligands in such important region of the active site.

The structure as described above was fully minimized with the sander program of Amber9, 200ps MD were performed on the hydrogens (all-atom model) and keeping the heavy atoms of the protein fixed, and then the structure was re-minimized. All simulations were performed at distance dependant dielectric constant $\varepsilon=4r$.

5.3 Rescoring Procedure

The rescoring procedure utilized here is developed by Prof. Giulio Rasteli and is based on the BEAR approach [216]. It is an automated procedure for the refinement and rescoring of virtual screening results. The procedure starts with an already prepared a pdb file of plasmepsin and docked conformations of the best 5000 compounds in mol2 file format. The procedure takes one ligand at a time, performs all the necessary steps required for one complete simulation, and takes the next ligand given in the input multi-mol file. In the first step, the procedure merges the co-ordinates of the docked conformation with the protein to create the complex. Then, antechamber is used to create a topology file of the ligand in which atoms are described with GAFF atom types and AM1-BCC charges. To make the procedure faster, atomic charges of the ligand are not computed during the procedure but read from the original mol2 file. This choice has the additional advantage that ligand charge calculations can be done once, and used for any target protein. Atomic partial charges of compounds in a database are calculated prior to the simulations by using antechamber. There is a possibility that Antechamber misses some force field parameters. The missing GAFF force-field parameters for the ligand are automatically assigned by parmcheck executable, and Amber topologies of ligand, receptor, and complex are created with leap (Amber 9).

Minimization, molecular dynamics and final re-minimization of the complexes are performed with distance-dependent dielectric constant 4ε, using the sander program. For each of these steps, the procedure is highly “flexible” in that it enables the user to set ad-hoc refinement options (for example, which residues are allowed to move during MD, the cutoff for non-bonded interactions, the number of cycles of minimization etc), depending on the application. After refinement of the complex, a pdb file is generated, and the final coordinates of the ligand, receptor and complex are updated, and used for binding free energy evaluation with Amber MM-PBSA and MM-GBSA. The free energy results (ΔG_{MM}, ΔG_{solv} and $\Delta G'_{bind}$) are written to a file, and the next compound in the database is analyzed.

The molecular dynamics simulation parameters used in the current study are minimization on the whole protein with distance-dependent dielectric constant ε=4r, with 2000 steps and a cutoff of 12A°. Molecular dynamics simulation is performed on the ligand alone at 300 K for 100 ps, with SHAKE turned on for bonds involving hydrogen, allowing a time-step of 2.0 fs.

5.4 Results

After rescoring the 5000 best docking results by molecular dynamics with Amber and MM-PBSA and MM-GBSA, the next step is to select the best compounds to test in the experimental laboratories. The output of the molecular dynamics simulation and rescoring are:

a. protein-ligand (P-L) complex after molecular dynamics step

b. P-L complex after MD and re-minimization step

c. A text file containing MM-PBSA and MM-GBSA scores.

A two-step criterion is employed for the final selection of the compounds. In the first step, selection of compounds is done based on the MM-PBSA and MM-GBSA scores.

The starting point of the analysis step is sorting the list of compounds according to MM-PBSA and MM-GBSA. Two independent lists of compounds were prepared, this is because both MM-PBSA and MM-GBSA are reliable in scoring. The reason, why one can rely on MM-GBSA and MM-PBSA is, due to the enrichment obtained on the known actives. Figure 30 and 31 demonstrate the retrieval of all the three known actives in the top 10 rank list, by both scoring functions. The MM-PBSA and MM-PBSA scores of the known actives are highlighted in Figure 30 and Figure 31. In the next step, top scoring 90 compounds from each list were analyzed manually. Each complex is visualized manually in 3D using the UCSF Chimera software. Visualization step is done, in order to find whether the compounds are making interactions to key residues of the protein and further to check the binding mode the compound. Figure 32 illustrates how visualization is performed.

The major criteria for selection are, the ligand making interactions to the two catalytic residues: ASH 34 and ASP214. Secondly, the interaction with other key amino acids: VAL78, SER79, SER218, PHE294, TYR192. The selection of key residues is based on the ligand plot information of the plasmepsin crystal structure. Figure 32 demonstrates the comparison of refined compounds to that of the co-crystallized ligand of 1LEE, R360.

The parameters used with Chimera are, intra model hydrogen bonds with relax constraints by 2.5 Angstrom and 20.0 degrees. The complexes without any interaction to one of the two amino acids of catalytic dyad (ASH34 and ASP214) are rejected and the complexes with at least one main interaction to amino acids of catalytic dyad were considered. In total 30 out of 180 compounds were selected and ordered from the vendor Chembridge (10 mg).

The final 30 compounds are listed in Table 14 along wih their MM-PBSA, MM-GBSA scores and the IC50 values. The compounds were identified as N-alkoxyamidine derivatives (Compounds 1-7 in the list), Guanidine derivatives (8-15), Amide derivatives (16-23), Urea and thiourea derivatives (24-29), Others (30).

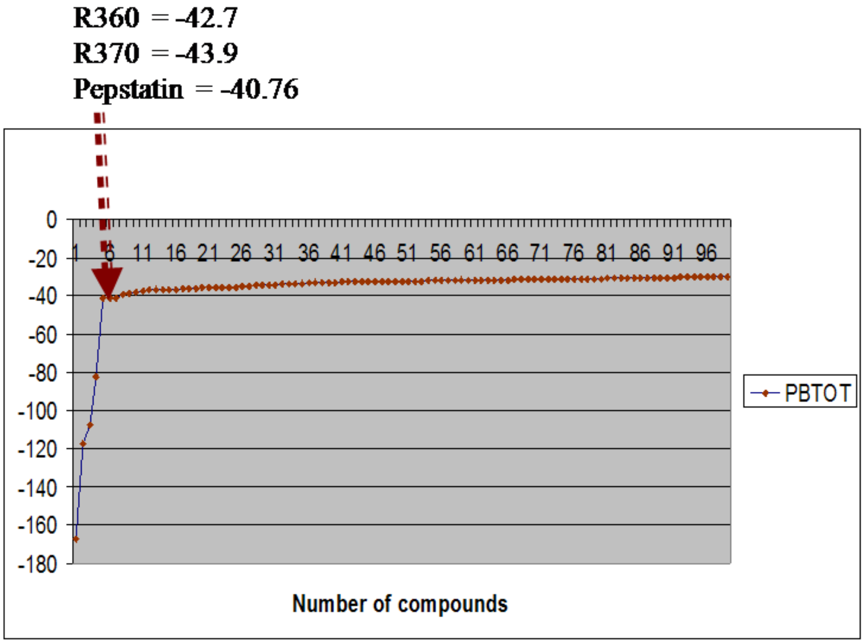


Figure 30: MM-PBSA scoring against plasmepsin docking conformations.
This Figure demonstrates the retrieval of known ligands (R360, R370, Pepstatin) in the top ten ranks according to MM-PBSA scoring function.

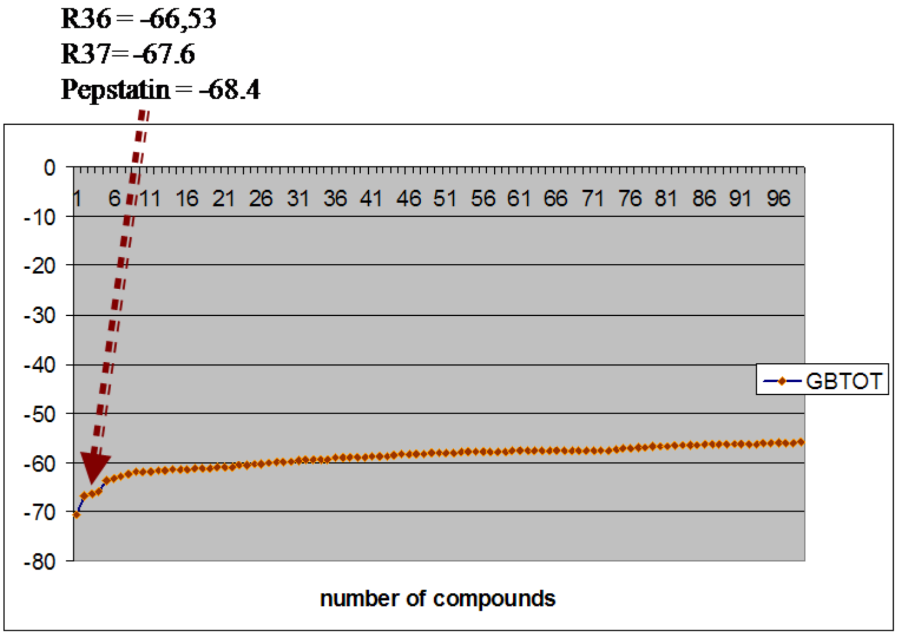


Figure 31: MM-GBSA scoring against plasmepsin docking conformations.
This Figure demonstrates the retrieval of known ligands (R360, R370, Pepstatin) in the top ten ranks according to MM-GBSA scoring function.

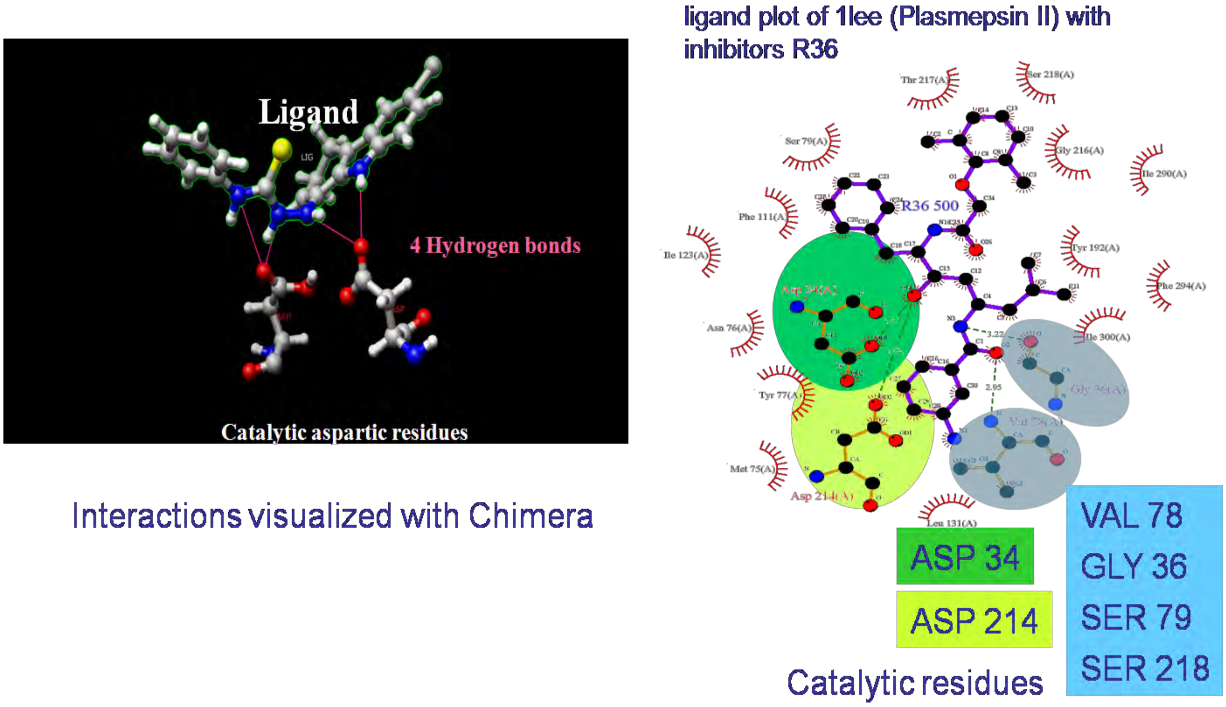
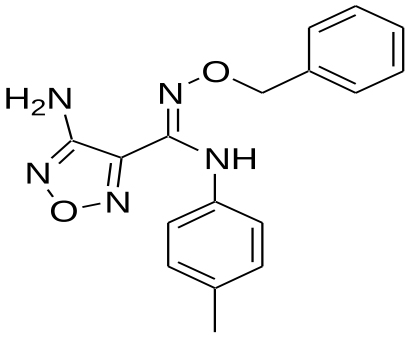
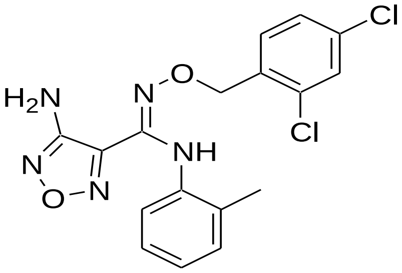
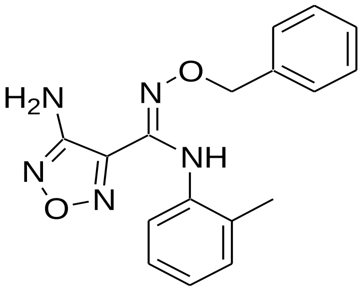
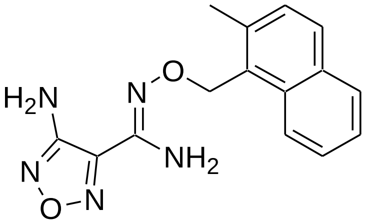
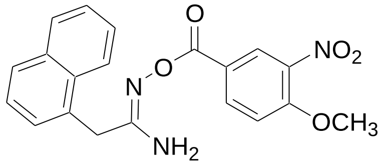
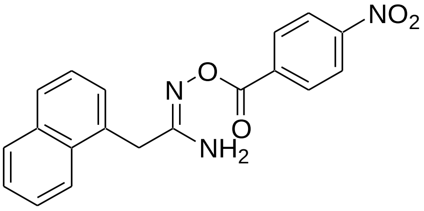
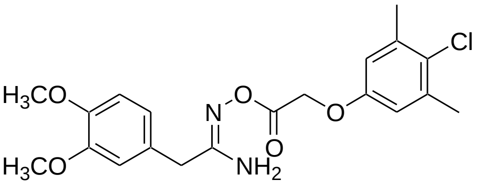
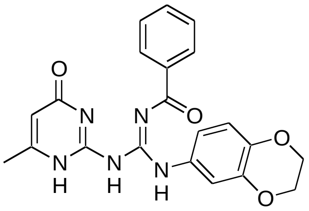
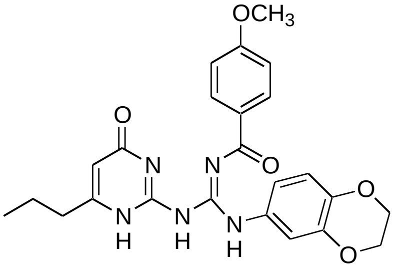
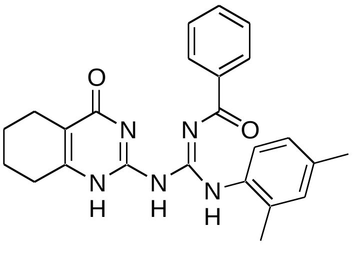
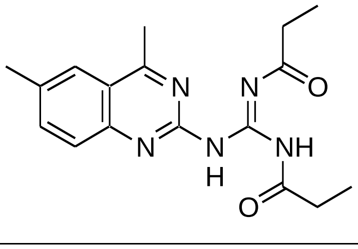
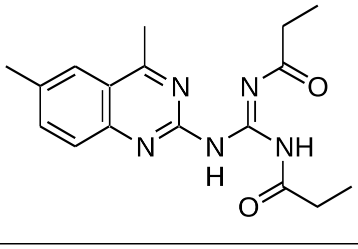
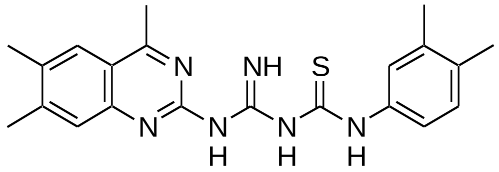
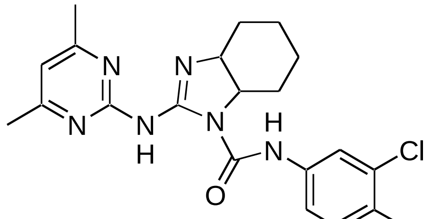
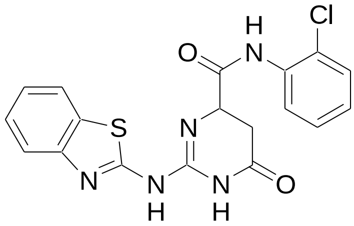
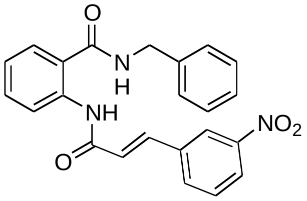
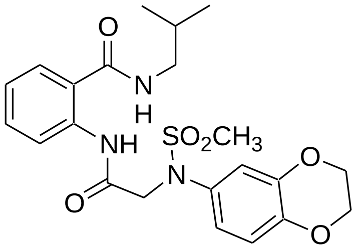
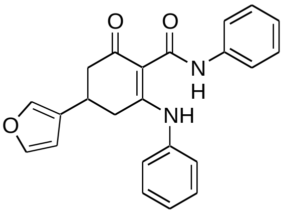
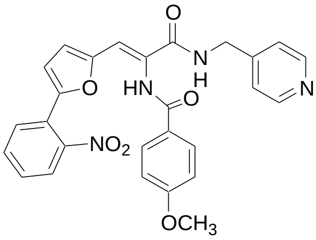
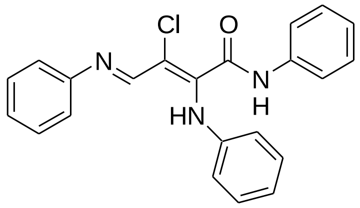
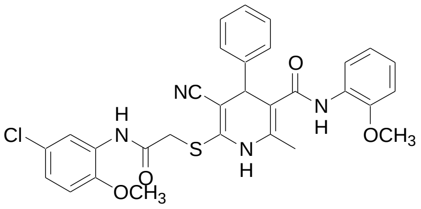
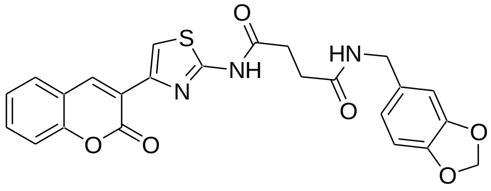
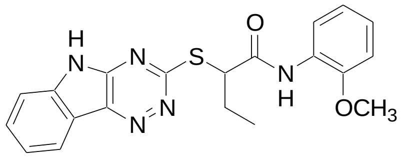
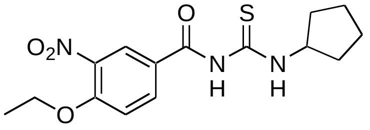
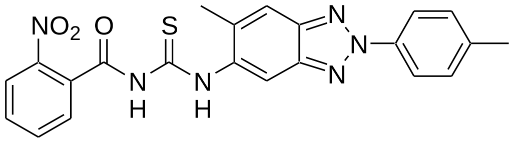
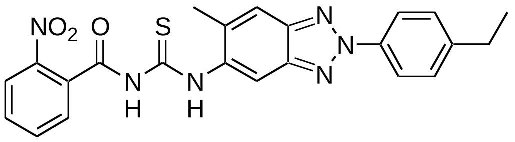
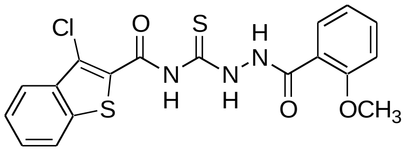
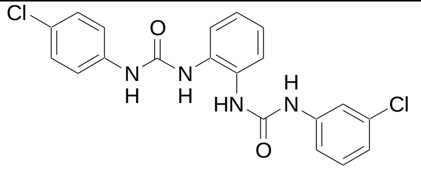


Figure 32: Analysis procedure employed for final selection of compounds.
The Figure on the left hand side demonstrate the chemical compound (compound after docking and refining by molecular dynamics simulations) making interactions to ASP34 and ASP214 (catalytic dyad), and these interactions are compared to the ligand plot information of 1LEE with co-crystallized ligand, R36, obtained from www.pdb.org (Right hand side). Only those compounds, which are similar R36 pose, are only considered for final testing.

S.No.	Molecule	MMPBSA score Kcal/Mol	MMGBSA score Kcal/Mol	IC50 values nM
1		-31.8	-54.2	305.1±1.5
2		-34.3	-63.3	5.5±2.0
3		-31.5	-54.8	6.4±0.7
4		-31.3	-46.3	42.6±1.5
5		-35.8	-54.2	236.4±0.7

6		-32.9	-55.4	145.2±2.4
7		-32.5	-61.1	4.3±0.6
8		-29.0	-66.4	62.1±0.6
9		-22.2	-59.4	118.1±1.9
10		-25.0	-61.5	8.8±0.8
11		-27.3	-57.9	n.i.
12		-41.3	-61.5	237.4±1.5

13		-21.5	-57.6	1087.6±0.7
14		-34.1	-66.9	9.5±1.1
15		-30.8	-54.3	96.1±0.2
16		-31.6	-57.6	30.0±1.8
17		-33.6	-58.0	n.i.
18		-31.7	-49.7	187.1±3.1
19		-23.8	-56.9	n.i.

20		-30.0	-50.0	189.0±1.4
21		-32.3	-55.2	57.3±0.4
22		-20.2	-55.5	n.i.
23		-32.7	-51.6	87.5±0.1
24		-30.9	-46.7	4.4±0.8
25		-20.4	-57.8	122.9±1.1
26		-29.0	-55.2	146.4±1.0
27		-32.5	-50.8	201.1±1.3
28		-27.5	-56.2	7.6±1.1

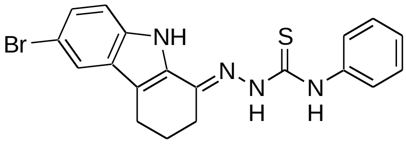
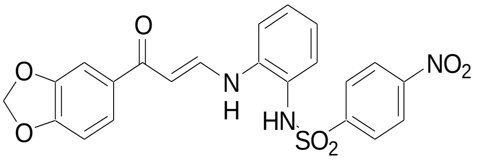
29		-31.6	-47.4	1831.3±1.9
30		-32.8	-55.9	38.9±2.4
I	**RS367**	-42.6	-66.7	18
II	**RS370**	-43.9	-67.6	30
III	**Pep.A**	-40.7	-68.4	4.3±0.9

Table 14: Final selection of compounds identified as plasmepsin inhibitors
Compounds 1-7 are N-alkoxyamidine derivatives, Compounds 8-15 are Guanidine derivatives, Compounds 16-23 are Amide derivatives, Compounds 24-29 are Urea and thiourea derivatives, Compound 30 is other compound. The IC50 values for RS367 and RS370 were taken from [208].

Diversity analysis

Diversity analysis was performed on the final 30 selected compounds to demonstrate that these compounds were diverse and contains several scaffolds (scaffold hopping). Lead hopping or scaffold hopping is defined as a technique that replaces the core part of the bioactive compound whilst retaining the bioactivity and the interactions made by the molecular fragments of the parent compound. The potential application of lead hopping are: overcoming the patent and IP issues (replacing the patented core scaffold with new core scaffold) and development of backup of series of compounds, i.e., even if one of the scaffold fails in the drug development, there is a possibility of replacement by other scaffold.

Fingerprints for all compounds were created by using FP: BIT MACCS, and then used Tanimoto coefficient (TC) as a similarity metric (TC) for calculating the diversity [223]. At similarity cut-off of Tanimoto coefficient 0.7, the 30 compounds were classified into different clusters, which indicates the final 30 compounds are diverse and dissimilar. Prior to the diversity test, 30 compounds were clustered into four classes manually. Through diversity test, it is proven that, even inside each cluster, the compounds are diverse, and thus are likely to enable lead hopping. Except for one cluster where four compounds are similar (big red box shown in Figure 33). The similarity is shown from green to red in the in Figure 33. With green stands for absolute dissimilar and red stands for absolute similar compounds.

The Tanimoto coefficient (similarity = the number of bits set in both molecules divided by the number of bits set in either molecule) is a validated and most commonly used similarity coefficient in chemical informatics while calculating diversity of the chemical compound database. It ranges from values 0 to 1, while value “1” corresponds to completely similar compound and “0” completely dissimilar).

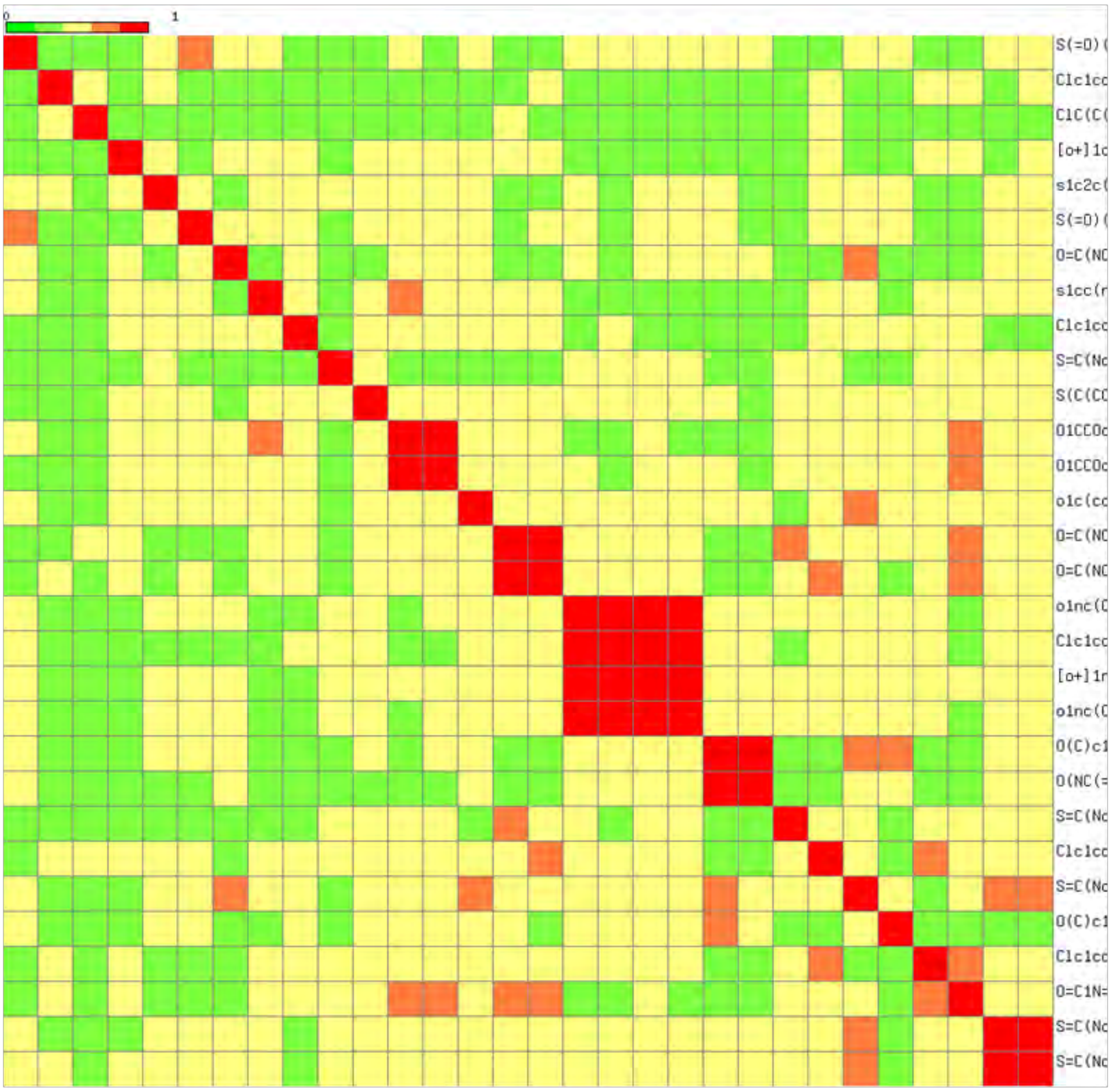


Figure 33: Diversity analysis of best 30 compounds against plasmepsin.
FP: BIT MACCS fingerprints and Tanimoto coefficient (TC) were used to calculate the diversity among the final 30 compounds selected. At similarity cut-off of TC 0.7, the 30 compounds were classified into different clusters. The similarity is shown from green to red. With green stands for absolute dissimilar and red stands for absolute similar compounds.

5.4.1 Experimental results

The experimental biochemistry assay results reported here are completely performed by Prof. Doman Kim and his colleagues, at Chonnam National University, South Korea. I include these results to illustrate that the in silico approach described in my thesis finally lead to biochemically active inhibitors.

Expression and preparation of recombinant plasmepsin: The gene of plasmepsin II [224, 225] was provided by MR4/American Type Culture Collection (USA) and expressed as inclusion bodies in *Escherichia coli* BL21(DE3)pLysS harboring PMII-pET3d. Expression and purification of recombinant plasmepsin II protein was conducted according to the method described by Hill et al [226]. with minor modification. Bacteria were grown by shaking in 1 litre of LB_{AMP} at 37 °C and the recombinant protein was induced by the addition of 400 µM isopropyl-β-D-thiogalactopyranoside (Biobasic, Korea). After the incubation with vigorous shaking at 16 °C for 18 h. *E. coli* were centrifuged and resuspended in lysis buffer (25 ml of 50 mM Tris-Hcl, 25 mM NaCl, pH 8.0) with 50 µl of β-mercaptoethanol (BME). The suspension was sonicated on ice to lyse the bacteria and centrifuged. The pellet containing inclusion bodies was washed in 25 ml of lysis buffer with 50 µl of BME, centrifuged, washed again with 0.1 M Tris pH 10 with 50 µl of BME and re-centrifuged. The pellet was re-suspended in 10 ml of 8 M urea in 100 mM Tris pH 8.0, 1 mM glycine, 1 mM EDTA and sonicated, after which 35 µl of BME were added. The suspension was stored at 4 °C overnight, and centrifuged at 23,000 x g for 30 min at 4°C. The supernatant was diluted 1:10 (v/v) in water and stirred overnight to allow refolding of recombinant protein. The refolded proteins were purified on a 50 ml Q-Sepharose Fast Flow (GE Healthcare, USA) equilibrated in 100 mM Tris-HCl (pH 8.5). After extensive washing, the recombinant protein was eluted with linear gradient of 0 - 1 M NaCl in the same buffer. The fraction containing the recombinant protein was concentrated and dialyzed in 10 mM Tris-HCl (pH 8.5), 5 mM NaCl, 20 mM BME. The purified protein was stored at -20 °C before using it for any assay.

FRET Substrate degradation inhibition assay: The substrate used for biological assay is a synthetic peptide (DABCYL-Glu-Arg-Nle-Phe-Leu-Ser-Phe-Pro-EDANS; Bachem, USA) designed to mimic the hemoglobin cleavage site. The substrate is conjugated with the fluorescent donor EDANS and the quencher DABCYL [227].The final volume of assay mixture is 50 µl. 5 µl of recombinant plasmepsin II (15 ng/µl) were acidified in 35 µl of assay buffer (100 mM sodium acetate, pH 4.5, 10% glycerol and 0.01% Tween 20). Inhibitors were dissolved in DMSO, serially diluted to the working concentrations (ranging from 1 nM to 10 µM) and 5 µl aliquots were added to the acidified enzyme. The final DMSO concentration in assay mixture was 1%. The enzyme and inhibitor mixture were incubated for 30 min at 37 °C and finally, incubated for 10 min at 37 °C with 5 µl of FRET substrate 50 µM, after that the fluorescence intensity (excitation 405 nm, emission 510 nm) was measured using a fluorescence plate reader SoftMax Pro 5 (Molecular Devices, USA).

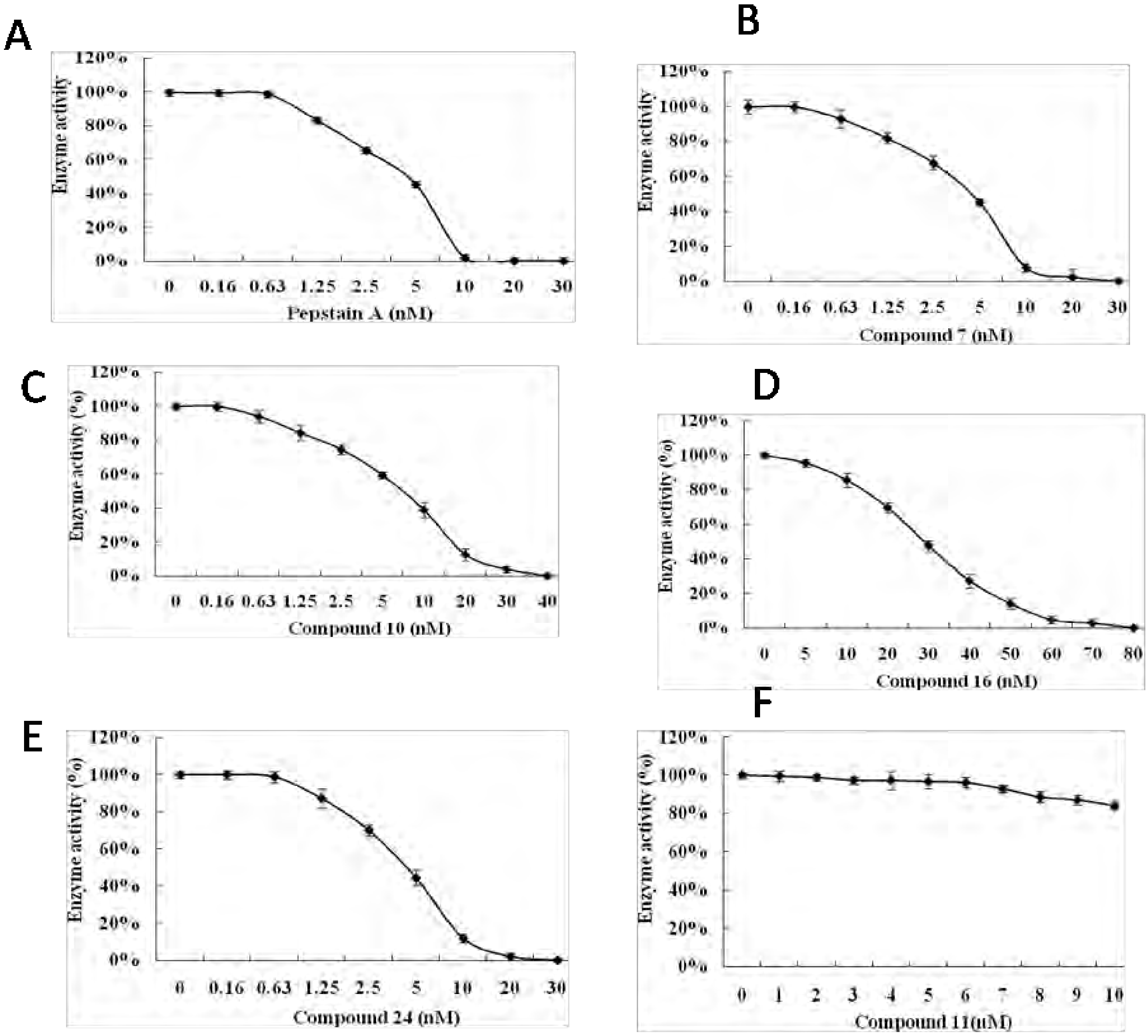


Figure 34: IC50 plots of five finally selected compounds and a control.
IC50 plots of: a) pepstatin A, b) compound 7, c) compound 10, d) compound 16, e) compound 24, and f) compound 11. Error bars represent standard deviations of the results from three independent experiments.

Biological testing

The 30 compounds in Table 14 are commercially available, and were purchased from ChemBridge at a purity grade higher than 95%. The compounds were tested against recombinant *P. falciparum* plasmepsin II using a well documented inhibition assay based on FRET substrate degradation [225, 227]. The known plasmepsin II inhibitor Pepstatin A was chosen as positive control [199]. A sample containing the assay mixture with 1% DMSO and without inhibitor was used as negative control. The IC50 values of the thirty compounds and

Pepstatin A are reported in Table 14. The Table 14 also reports IC50 data of RS367 and RS370 taken from the literature [208]. Remarkably, 26 compounds were identified as active inhibitors with IC50 values ranging from 4.3 nM (compound 7) to 1.8 µM (Compound 29), while four compounds (11, 17, 19, 22) were inactive. Interestingly, seven compounds (2, 3, 7, 10, 14, 24, 28) showed IC50 values similar to Pepstatin A (IC50 of 4.3 nM). Figure 34 reports the IC50 plots for Pepstatin A (Figure 34A), four of the most active compounds (7, 10, 16 and 24) taken as representatives of each chemical class here investigated (Figure 34B-E), and the inactive compound 11 (Figure 34F).

Chemical compounds: All the thirty tested compounds were commercially available and provided by ChemBridge. Even though, the purity grade of the supplier certificated compounds higher than 95%, CNH combustion analysis was performed to evaluate their purity. Rossella Gallesi, University of Modena has kindly performed the combustion analysis. All the 30 compounds showed purity higher than 95%.

5.5 Summary

The application of refinement and rescoring procedure for post-docking analysis and selection of molecules for biological testing, led to the selection of 30 compounds with favorable predicted binding free energies and interaction with key plasmepsin residues, belonging to four different chemical classes. The impact of molecular dynamics simulations followed by MM-PBSA and MM-GBSA is quite significant, while selecting the compounds for experimental testing. The in vitro assay revealed that 26 of the 30 compounds selected were able to inhibit plasmepsin II with IC50 values in the range of 4 nM to 2 mM. These results are very encouraging and suggest that the overall approach (docking as first step followed by molecular dynamics simulations and rescoring by MM-PBSA and MM-GBSA) used to select the candidate molecules can be used to discover new plasmepsin inhibitors. Six out of the 26 compounds exhibited better activity than very-well known protease inhibitor, pepstatin A. Notably, among the chemical classes discovered in this study, only urea compounds have been previously reported to be plasmepsin II inhibitors (so-called “Walter Reed” compounds).The remaining classes effectively provide novel and interesting opportunities for developing compounds with potential antimalarial activity. These compounds are currently under evaluation for their inhibitory activity on parasite growth and for their potential toxicity on human cells. In addition, further studies will be undertaken to investigate the SAR of these new inhibitors.

6 Chapter 6: Large-scale Virtual screening on multiple targets of malaria

Currently existing antimalarial drugs are targeting single target. However, due to the complexity of the Plasmodium life cycle and drug resistance more targets and more metabolic pathways have to be targeted to counteract the disease. In other words, for the effective treatment of the disease, it is necessary to identify drugs that have novel mechanism of action and are further able to target multiple targets at the same time. Moreover, it will be a benefit, if the newly identified drugs target different stages of plasmodium life cycle or target proteins/enzymes involved in different metabolic pathways. This multi-target approach will definitely overcome drug resistance, which is a major problem haunting antimalarial drug discovery.

Followed by the success achieved on the virtual screening plasmepsin, (WISDOM-I reported in chapter 4) both on the computation and biological sides, several scientific groups around the world proposed targets, which led to the second assault on malaria, i.e., WISDOM-II. The WISDOM-II project deals with several targets implicated in malaria (mostly X-ray crystal structures). Targets from different classes of proteins are tested; reductases such as malarial dihydrofolate reductase (DHFR) and transferases such as glutathione-S-transferase (GST), see Table 15. Hence, in the current chapter: WISDOM-II[4], proteins that are involved in different metabolic activities of the parasite are selected.

As different species of Plasmodium cause malaria to humans, proteins not only from *Plasmodium falciparum,* but also from the *Plasmodium vivax* are embattled. The extension of the work to target *P. vivax* is, due to its resurgence and casualties caused [228]. From Table 15, it is clear that targets are chosen as such to identify novel inhibitors for different proteins implicated in malarial life cycle with the idea in mind to interfere with resistance.

This chapter reports the large-scale virtual screening effort on the multiple targets, with focus on the improvements to the existing workflow described in chapter 4, such as novel procedures and strategies for storage, post-processing, analysis of the docking results, and finally selecting a representative set of potential inhibitors for further *in vitro and in vivo* testing. The main goal of the WISDOM-II project is to identify broad range of inhibitors that are active against multiple independent targets implicated in malaria.

[4] This chapter is based on Kasam V et al., WISDOM-II: Screening against multiple targets implicated in malaria using computational Grid infrastructures. Malaria Journal, 2009, 8:88.

Target	Activity	Structure	PDB id	Resolution Å	Cocrystallized Ligand	Co-factor
*Pf*GST	Detoxification	Dimer	1Q4J	2.2	GTX	NO
Pf DHFR (wild type)	DNA synthesis	Polymer	1J3I	2.33	WR99210	NADPH
Pf DHFR (Quadruple mutant)	DNA synthesis	Polymer	1J3K	2.10	WR99210	NADPH
*Pv*DHFR (wild type)	DNA synthesis	Polymer	2BL9	1.90	Pyrimethamine	NADPH
*P*vDHFR (Double mutant)	DNA synthesis	Polymer	2BLC	2.25	Des-chloropyrimethamine	NADPH

Table 15: Structual features of potential targets identified for the WISDOM-II project.

6.1 Target structures

6.1.1 Glutathione-S-transferase.

The *P. falciparum* glutathione S-transferase enzyme belongs to a super family of multifunctional, dimeric, phase II detoxification enzymes that can bind various xenobiotic, electrophilic substrates. Parasites as well as other rapidly dividing cells are highly dependent on a functional antioxidant defense system. For most parasites the sources of reactive oxygen species is mainly their high metabolic rate as well as oxidative stress imposed by the host's immune system. Additionally, the *P. falciparum* parasite performs haemoglobin degradation - a source of oxidative stress and free radicals [229]. The antioxidant defense system of *P. falciparum* is therefore a pathogenicity mechanism; an ensemble of antioxidants like glutathione as well as antioxidant enzymes mediates it [229].

The primary function of GST lies in the protection of cellular macromolecules. GST deactivates harmful chemicals via the nucleophilic addition of the thiol (SH) group from glutathione (GSH), to the hydrophilic moiety of the toxic agent, thus rendering the electrophilic compounds harmless and enabling the removal of the substance. Because of the inactivation of potentially hazardous substances, GST activity is beneficial to an organism's health and survival [230, 231]. In chloroquine-resistant parasites, GST activity is directly related to drug pressure [232, 233].

Inhibition of GST will impair the general detoxification processes and, because the enzyme has peroxidase activity, reduce the antioxidant capacity of the parasite [234].

PfGST (EC 2.5.1.18) is a multi-functional protein consisting of two monomers. In accordance with other GST enzymes each monomer of PfGST contains an N-terminal α/β domain and C-terminal α-helical domain. The active site is defined by two binding sites: the G site, which binds GSH, and the more flexible H site, which can bind various other substrates. Hydrophobic effects predominantly hold the monomers together, but four salt bridges and four hydrogen-bonded pairs of residues also contribute to the dimerization [235, 236]. The G site is relatively rigid and not greatly affected by inhibitor binding, with the exception of the C-terminal tail and the loop connecting the α-4 and α-5 helices. This region is very specific for its natural substrate (GSH). The recognition and binding occur via a network of polar interactions between PfGST and GSH.

The hydrophobic binding pocket (H site) is considerably more variable than the G site, due to the nature of second substrates. The substrate specificity of different isozymes in the GST super family can be attributed to the variation of amino acids present in the H site consequently leading to different interactions a ligand can form with amino acids in the H site of the enzymes [237].

PfGST also possesses a short μ-loop. In contrast to other μ-class GST enzymes, PfGST has only five residues after α-8, which is too short to form a wall or α-helix. This feature is lacking in PfGST, resulting in a more solvent-accessible H site. The result is that the H site is less shielded from solvents [237, 238].

6.1.2 *Plasmodium vivax* and *Plasmodium falciparum* DHFR.

Plasmodium vivax is becoming resistant to chloroquine and other antifolates, such as pyrimethamine [239, 240, 241]. The target enzyme of pyrimethamine is dihydrofolate reductase (DHFR). It was demonstrated that the resistance to pyrimethamine is caused by a point mutation [242]. Interestingly, the crystal structure of DHFR enzyme from *P. vivax* was published by Kongsaeree *et al* in 2005 [243], where they indicated that the principal difference between DHFR wild type and mutant, implicated in the antifolate resistance, is a structural change in the chain of Asn108, and this steric conflict is not present in *P. falciparum*.

Antifolates, such as pyrimethamine and cycloguanil, are the most exploited class of anti-malarials. To date, the most widely used antifolate is a combination of pyrimethamine, a dihydrofolate reductase (DHFR) inhibitor, and sulphadoxine, a dihydropteroate synthase (DHPS) inhibitor. DHFR and DHPS are two enzymes that belong to the folate biosynthetic pathway [19]. Although their synergistic action results in enhanced activity, drug resistance seriously compromises their efficacy. As a major advance towards the understanding of drug resistance in malaria, it has been demonstrated that drug resistance is due to single and multiple mutations of various amino acids in the DHFR and DHPS active sites in *P. vivax* as well as *P. falciparum* [244, 245]. The analysis of the gene encoding *P. falciparum* DHFR from resistant parasites suggested that antifolate resistance arises from point mutations in the DHFR domain, mainly at positions 16, 51, 59, 108, and 164. It has been demonstrated that parasites with mutations at 16 and 108 have developed resistance to cycloguanil, with a thousand-fold drop in the binding affinity (K_i) compared with the wild type, whereas the K_i of pyrimethamine is almost unaffected. On the contrary, there is cross-resistance between the drugs when multiple mutations at position 51, 59, 108 and 164 occur.

Combined homology modeling and molecular dynamics simulation studies proposed how pyrimethamine, cycloguanil and WR99210 (a third-generation antifolate) bind to wild type and resistant mutant *P. falciparum* and *P. vivax* DHFRs [246, 247]. Crystal structure determination of the malarial DHFRs in complex with antifolates have confirmed and strengthened the proposed binding modes [248, 247].

6.2 Virtual docking procedure

The different steps of the virtual docking procedure will be described in the following section.

6.2.1 Target preparation

Glutathione-S-transferase (GST)

The X-ray crystal structure of GST utilized is 1Q4J [33]. 1Q4J is a homo-dimer with two chains: A and B. In the first step, all the crystal water molecules were removed from the protein. The active site is defined as 8Ǻ around the co-crystallized ligand: GTX. All the residues that are significant for activity of the protein and binding of the ligand are included in the active site. Re-docking with GTX ligand is performed for further optimization of the target parameters as well as software parameters.

***Plasmodium vivax* DHFR (PvDHFR) and *Plasmodium falciparum* DHFR (PfDHFR).**

The protein structures used in this investigation are the crystal structures of wild-type *P. falciparum* DHFR (PDB code 1J3I), and of its N51I+C59R+S108N+I164L highly resistant mutant (PDB code 1J3K), both in complex with NADPH and the potent inhibitor WR99210. The structures of wild type *P. vivax* DHFR (PDB code 2BL9) and of its S58R+S117N resistant mutant are (PDB code 2BLC) in complex with pyrimethamine and des-chloro pyrimethamine, respectively [243, 247]. The structures were cut at residue Asn231, which corresponds to the DHFR domain of the bifunctional DHFR-TS structure. Of the dimer, unit B was chosen because of its less missing residues. Met1 was built as in unit A, and the position of missing residues from Asp87 to Asn90 was modeled using Modeller software [249]. At this purpose, the enzyme sequence with the four missing residues was aligned with the complete sequence, and ten models were generated with the Modeller software using 1J3I as template. The best model according to Prosa II was saved, and the coordinates of the four missing residues were inserted back in the original crystal structure. For the quadruple mutant of *P. falciparum* DHFR, the missing segment from residue 81 to 97 was taken from the wild type structure. *P. vivax* DHFR was prepared using the same methodology. Residues E24 and K48, which have truncated side chains in the original crystal structures, were assigned based on standard Amber topologies of amino acids. Residues from 84 to 105 missing in the double mutant structure were taken from the wild type structure.

All water molecules in the crystal structure were removed except for two conserved waters embedded into the protein (corresponding to W1249 and W1250 in the original 1J3I crystal structure) and close to the important residue D54. Hydrogens were added to the structures using the internal coordinates of the AMBER all-atom database. All Lys and Arg residues were positively charged and Glu and Asp residues negatively charged. All calculations were performed with AMBER9 and the ff03 force field [250]. The parameters of the cofactor NADPH were taken from previous simulations [250, 248].

The structures prepared as described above were refined with energy minimization, employing a distance-dependent dielectric constant e=4r and a cutoff of 12Å for non-bonded interactions. Firstly, 500 steps of conjugate gradient energy minimization were performed on the hydrogen atoms only, followed by 5,000 steps of minimization on the entire structure. Then, in order to refine the position of the hydrogen atoms added with Amber, 50ps molecular dynamics at 300°K was performed on the hydrogens by adding strong restraints on the heavy atoms. Finally, 5000 steps of minimization were performed without restraints. All minimizations were performed on the protein structures with the corresponding antifolates bound in the active site (WR99210 or Pyr). For the antifolates, partial atomic charges on atoms were calculated with the AM1-BCC method [251] implemented in the antechAmber module of Amber9. Atom types and missing force-field parameters of the ligands were assigned based on the General Amber force-field (gaff) [252].

The target structures, *P. falciparum* and *P. vivax* DHFR were proposed by Prof. Guilio Rastelli, University of Modena, Italy. The initial preparation of the target and analysis of the final results were also performed by Prof. Guilio Rastelli group.

6.2.2 Setting up the platform before large-scale virtual screening

Docking was performed on ZINC database, which is collection of 4.3 million ready to dock chemical compounds. The docking software utilized is FlexX. FlexX [209, 78] and ZINC database are already discussed in chapter 4. The parameter sets that were utilized in these experiments are identical to those described in chapter 4 (Maximum overlap volume and Place particles).

Re-docking

The results of the re-docking experiments are displayed in Table 16. Results are analyzed at three levels: the RMSD (root mean square deviation) between the docking pose and the co-crystallized ligand, the docking score and the interaction information between protein and ligand. The docking poses of the co-crystallized ligands generated by FlexX are manually visualized, and compared to their respective ligand plots. Two aspects were considered; the binding mode of the docking pose should be similar to ligand plot and should make interactions to the key residues of the receptor as described in ligand plots (www.pdb.org). Table 16 displays the docking score and RMSD of the best docking conformation.

Target	Ligand	Total Score	RMS-Value	Total Score	RMS-Value	Total Score	RMS-Value	Total Score	RMS-Value
			1		2		3		4
1Q4J_a	GTX	-24.33	3.68	-20.99	7.53	-20.15	6.94	-25.93	7.11
1Q4J_b	GTX	-19.93	6.45	-18.33	11.83	-18.33	11.78	-25.07	6.28
2BLC	CP7	-13.47	4.88	-14.09	4.78	-12.53	4.45	-14.43	4.78
2BL9	CP6	-13.657	4.71	-12.50	6.17	-13.65	4.71	-12.50	6.17
Target	**Ligand**	**Total Score**	**RMS-Value**	**Total Score**	**RMS-Value**	**Total Score**	**RMS-Value**	**Total Score**	**RMS-Value**
			5		**6**		**7**		**8**
1J3K	WR9	-24.33	3.68	-23.91	1.81	-23.75	2.21	-26.41	3.14
1J3I	WR9	-30.75	2.49	-21.98	1.41	-20.83	2.69	-25.69	1.76

Table 16: Re-docking results of different targets in different parameter sets of FlexX

For target PfGST (1Q4J: PDB ID), parameter set 1 performed better compared to other parameter sets. However 3.68 Å is still a big deviation (ideal RMSD should be <2Å), but the binding mode the co-crystallized ligand adopted was quite convincing, as the docking pose was making interactions to the key residue. Besides that, the docking pose made interactions to the key residues responsible for the activity of the protein. In case of *P. vivax* DHFR (2BLC and 2BL9), the docking of the co-crystallized ligand did not perform well. The RMSD deviations were high (>4Å) and the binding modes were not convincing. This is due to clashes between the protein and ligand atom surfaces. For PfDHFR, re-docking was performed against protein structures before and after minimization by Amber software. Docking software parameters were tuned accordingly. For PfDHFR, we increased the maximum allowed overlap between the protein and ligand atom to diminish the van der Waals clashes. Re-docking against minimized structures with the same parameter sets gave best results. All the parameter sets reproduced the actual binding mode of the ligand, further made interactions to key amino acids and RMS deviation were less than 2 Å. Re-docking results of PfDHFR minimized structures are displayed in Table 16. Besides docking the co-crystallized ligand, well-known inhibitors against PfDHFR are docked. Table 17 and 18 displays the results of cycloguanil and pyrimethamine, WR9 under different docking parameter sets. Parameter 8 (maximum allowed overlap volume between protein and ligand surface:100Å^3)

gave the best results in terms of docking score, docking conformation and interactions to key amino acids.

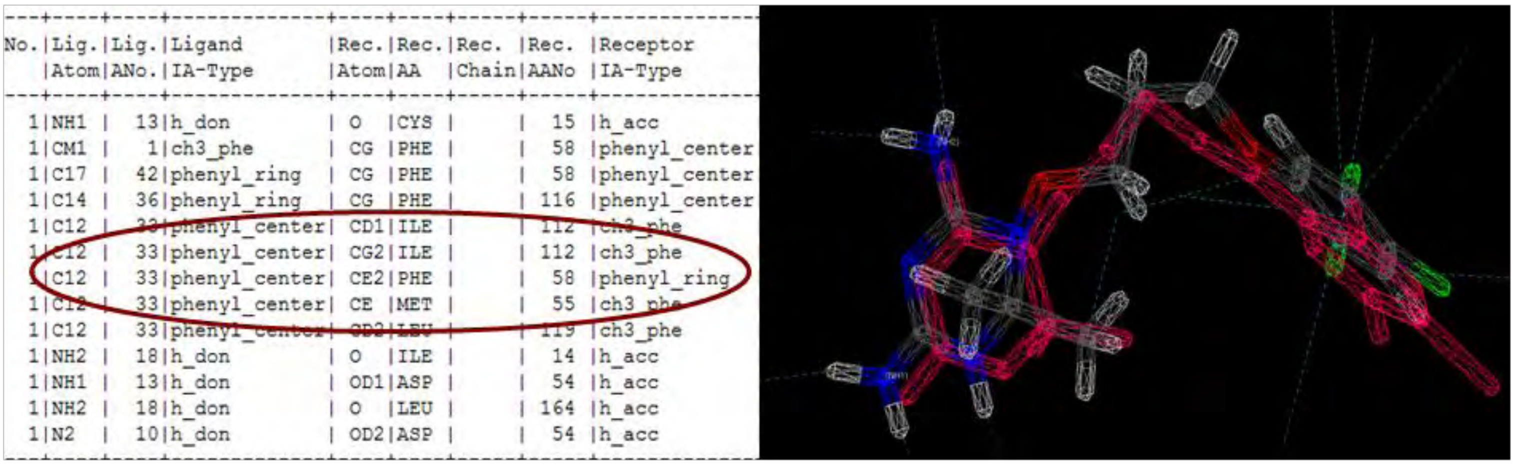
	Best Score (kJ/mol)	RMSD for best solution (Å)	Rank for Best RMSD solution	Score for best RMSD Solution	Best RMSD (Å)
QM_WR9_10	-23.22	3.37	106	-9.85	0.76
QM_WR9_20	-23.91	1.81	158	-9.85	0.76
QM_WR9_30	-23.75	2.21	97	-12.57	0.75
QM_WR9_100	**-26.41**	**3.14**	**40**	**-18.06**	**1.14**
QM_CYC_10	-23.25	1.46	525	-12.40	0.92
QM_CYC_20	-23.25	1.46	525	-12.40	0.92
QM_CYC_30	-23.25	1.46	146	-20.39	0.97
QM_CYC_100	**-23.25**	**1.46**	**699**	**-15.08**	**1.01**
QM_PYR_10	-23.68	1.21	8	-21.80	0.69
QM_PYR_20	-23.68	1.21	8	-21.80	0.69
QM_PYR_30	-23.60	1.26	16	-22.08	0.74
QM_PYR_100	**-21.95**	**1.51**	**20**	**-20.31**	**0.97**

Table 17: Re-docking results against quadrupule mutant DHFR.
QM= Quadruple mutant, WT = Wild type; RMSD is in Angstroms; Score is the free energy in kJ/mol

The results displayed in Table 17 correspond to the quadruple mutant results (1J3I: PDB ID) and Table 18 corresponds to the wild type results (1J3K: PDB ID). Figure 35 and 36 display the re-docking pose of WR9 against minimized structure of the Pf DHFR (1J3K) and PfDHFR (1J3I), respectively, on the right hand side of the Figure 35 and 36, we can see the docking pose (CPK color) and reference co-ordinates in red color (IJ3K) and violet color (1J3I). On the left hand side protein-ligand interactions are displayed. Highlighted are the interactions responsible for the activity of the protein (parameter sets 5, 6, 7, 8 correspond to maximum allowed overlap volume 10, 20, 30, 100 Å3 respectively).

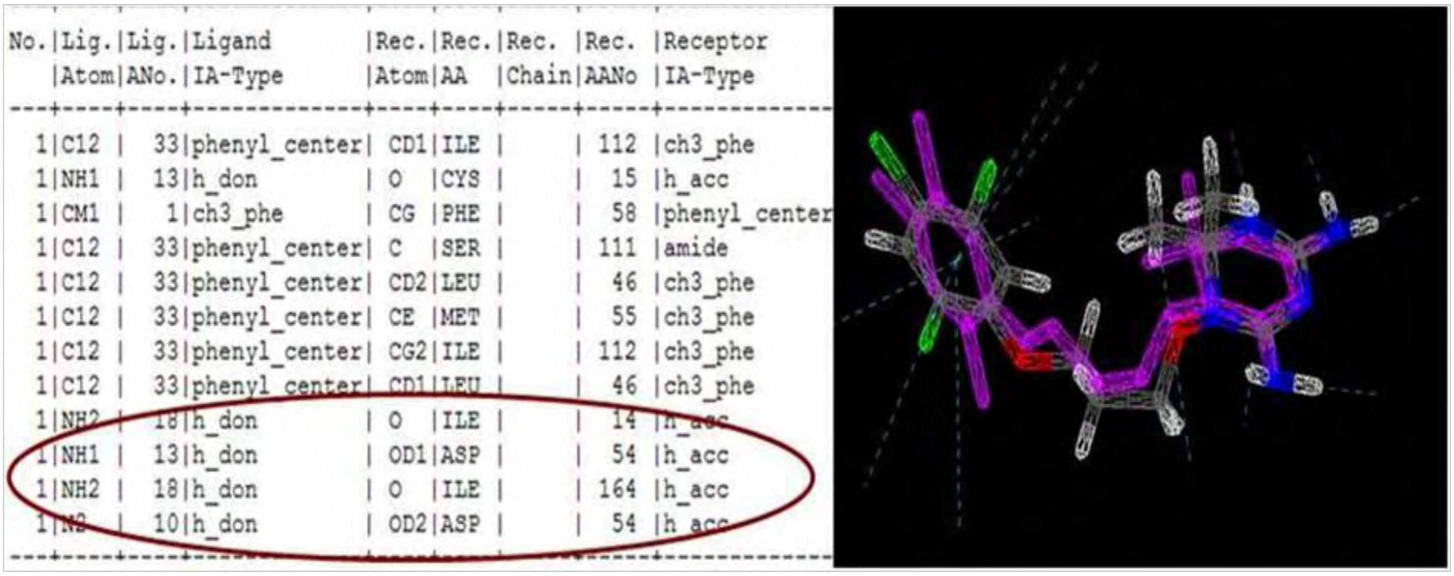
	Best Score (kJ/mol)	RMSD for best solution (Å)	Rank for Best RMSD solution	Score for best RMSD Solution	Best RMSD (Å)
WT_WR9_10	-30.75	2.49	57	-21.98	0.91
WT_WR9_20	-21.98	1.41	46	-13.78	0.91
WT_WR9_30	-20.83	2.69	2	-19.67	0.99
WT_WR9_100	**-25.69**	**1.76**	**4**	**-22.47**	**0.83**
WT_CYC_10	-24.36	1.43	622	-19.60	0.89
WT_CYC_20	-24.47	1.46	720	-19.88	0.95
WT_CYC_30	-24.47	1.46	7	-22.16	0.95
WT_CYC_100	**-24.70**	**1.49**	**11**	**-31.49**	**0.97**
WT_PYR_10	-29.72	1.25	6	-28.02	0.46
WT_PYR_20	-29.73	1.26	2	-27.70	0.53
WT_PYR_30	-29.73	1.26	2	-27.70	0.53

Table 18: Illustrates re-docking results against wild type DHFR.
QM= Quadruple mutant, WT = Wild type; RMSD is in Angstroms; Score is the free energy in kJ/mol



```
---+----+----+-------------+----+----+-----+-----+-------------
No.|Lig.|Lig.|Ligand       |Rec.|Rec.|Rec.  |Rec.  |Receptor
   |Atom|ANo.|IA-Type      |Atom|AA  |Chain|AANo |IA-Type
---+----+----+-------------+----+----+-----+-----+-------------
  1|NH1 |  13|h_don        | O  |CYS |     |  15 |h_acc
  1|CM1 |   1|ch3_phe      | CG |PHE |     |  58 |phenyl_center
  1|C17 |  42|phenyl_ring  | CG |PHE |     |  58 |phenyl_center
  1|C14 |  36|phenyl_ring  | CG |PHE |     | 116 |phenyl_center
  1|C12 |  33|phenyl_center| CD1|ILE |     | 112 |ch3_phe
  1|C12 |  33|phenyl_center| CG2|ILE |     | 112 |ch3_phe
  1|C12 |  33|phenyl_center| CE2|PHE |     |  58 |phenyl_ring
  1|C12 |  33|phenyl_center| CE |MET |     |  55 |ch3_phe
  1|C12 |  33|phenyl_center| OD2|LEU |     | 119 |ch3_phe
  1|NH2 |  18|h_don        | O  |ILE |     |  14 |h_acc
  1|NH1 |  13|h_don        | OD1|ASP |     |  54 |h_acc
  1|NH2 |  18|h_don        | O  |LEU |     | 164 |h_acc
  1|N2  |  10|h_don        | OD2|ASP |     |  54 |h_acc
---+----+----+-------------+----+----+-----+-----+-------------
```

Figure 35: Illustrates the re-docking of WR9 ligand against 1J3K in parameter 8



```
No.|Lig.|Lig.|Ligand        |Rec.|Rec.|Rec. |Rec. |Receptor
   |Atom|ANo.|IA-Type       |Atom|AA  |Chain|AANo |IA-Type
---+----+----+--------------+----+----+-----+-----+------------
  1|C12 |  33|phenyl_center| CD1|ILE |     | 112 |ch3_phe
  1|NH1 |  13|h_don         | O  |CYS |     |  15 |h_acc
  1|CM1 |   1|ch3_phe       | CG |PHE |     |  58 |phenyl_center
  1|C12 |  33|phenyl_center| C  |SER |     | 111 |amide
  1|C12 |  33|phenyl_center| CD2|LEU |     |  46 |ch3_phe
  1|C12 |  33|phenyl_center| CE |MET |     |  55 |ch3_phe
  1|C12 |  33|phenyl_center| CG2|ILE |     | 112 |ch3_phe
  1|C12 |  33|phenyl_center| CD1|LEU |     |  46 |ch3_phe
  1|NH2 |  18|h_don         | O  |ILE |     |  14 |h_acc
  1|NH1 |  13|h_don         | OD1|ASP |     |  54 |h_acc
  1|NH2 |  18|h_don         | O  |ILE |     | 164 |h_acc
  1|N2  |  10|h_don         | OD2|ASP |     |  54 |h_acc
---+----+----+--------------+----+----+-----+-----+------------
```

Figure 36: Illustrates the re-docking of WR9 ligand against 1J3I in parameter 8
On the left hand side of both the figures, Interaction information between ligand atom and target protein are displayed. On the right hand side redocked pose (CPK color) and reference coordinates (Red color) are displayed

Virtual screening on the EGEE Grid infrastructure

After setting up the docking platform, virtual screening was performed on 4.3 million compounds against the targets specified in Table 15. Screening 4.3 million compounds on multiple target structures is done on EGEE and its related Grid infrastructures (AuverGrid, EELA, EUChinaGrid and EUMedGrid). Deployment of docking jobs and WISDOM production environment is already discussed in chapter 3.

6.2.3 Database schema to store the results

The outputs of the docking results in FlexX are log files. All the results are stored and analysed by using MySQL databases. Three different forms of results are saved and analysed from each docking assay:

i. Docking scores of the ten best solutions after clustering

ii. Interaction information between protein and ligands of the ten best solutions,

iii. Binding modes of the ten best solutions.

During the first deployments (WISDOM-I-chapter 4) the results were stored on the Grid storage elements using the Grid data management, this format made the analysis of the results particularly difficult. Since, docking and scoring results often need to be extracted, parsed, and analyzed by biologists, user-friendly data retrieval systems are needed.

Hence, it was decided to store the docking results in a relational database. The relevant information (docking score and interaction information) were parsed directly into a relational database, and the user (biologist) can use SQL queries to find and retrieve relevant information.

The database is designed around the docking table (Figure 37). The total docking score in FlexX is the result of six different individual energy contributions. Along with the total score, these six individual energy contributions are also stored. The insertion of records was performed directly at the end of the docking jobs on the Grid. The raw result files (Docking results in log files) are also stored and replicated on the Grid storage elements for the backup.

A simple perl script, using perl DBI library, parses the result files, builds the useful information, and further insert the data from the Grid to a remote MySQL server.

The real interest of such a solution is that the useful data are immediately available for query and analysis during the process. The usage of relational database along with SQL eases the selection of the best compounds, as they can be selected accordingly to any attribute of the database tables. As almost all programming languages offer the ability to access database management systems through APIs, it will also ease the interoperability with web servers, for instance, if one wants to be able to monitor and view the data on a web interface.

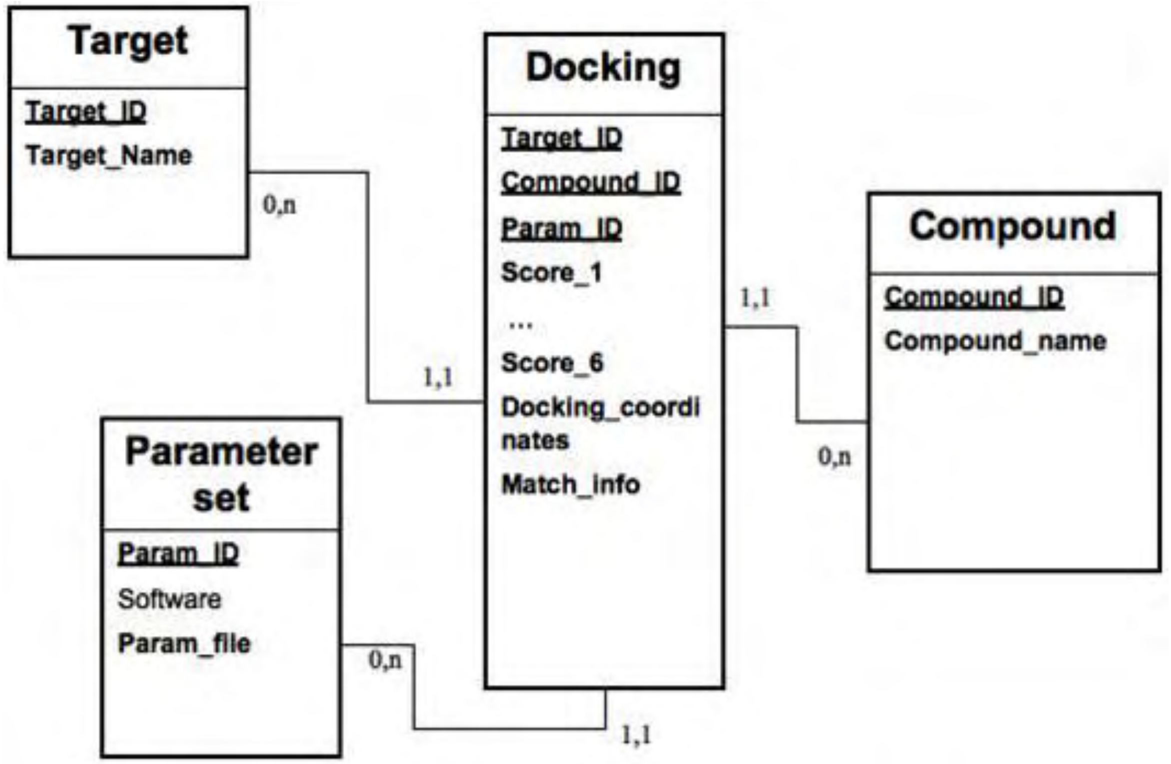


Figure 37: A view of the result database schema used to store and analyze docking results in WISDOM-II.

6.2.4 Strategies adopted for analysing the results

During WISDOM-I project, described in chapter 4, the compounds with best docking scores were visualized manually. Interestingly it was observed that, some of the top scoring compounds were making interactions to the key residues, but the binding modes of the compounds were not optimal and not comparable with the ligand plots of co-crystallized ligands, one of the reason may be due to the rigid nature of the receptor

Consequently, as described in chapter 5, the result analysis took place by first extracting the compounds based on docking scores and then by rescoring the best docked ligands with more sophisticated scoring functions. Such workflow, called BEAR (Binding Estimation After Refinement) [216] significantly improved the overall procedure and resulted in the identification of plasmepsin inhibitors (WISDOM-I).

Hence, the same workflow: Initial selection of compounds by docking score and further rescoring by BEAR procedure is utilized in WISDOM-II. The best 5,000 scoring conformation (against GST) are extracted and subjected them to molecular dynamics simulations and MM-PBSA and MM-GBSA calculations [219, 217]. After rescoring by molecular dynamics methods, the compounds are further manually visualized by using Chimera software [253] and other structural visualizing software. Figure 38 illustrates the complete filtering process employed in the WISDOM-II project.

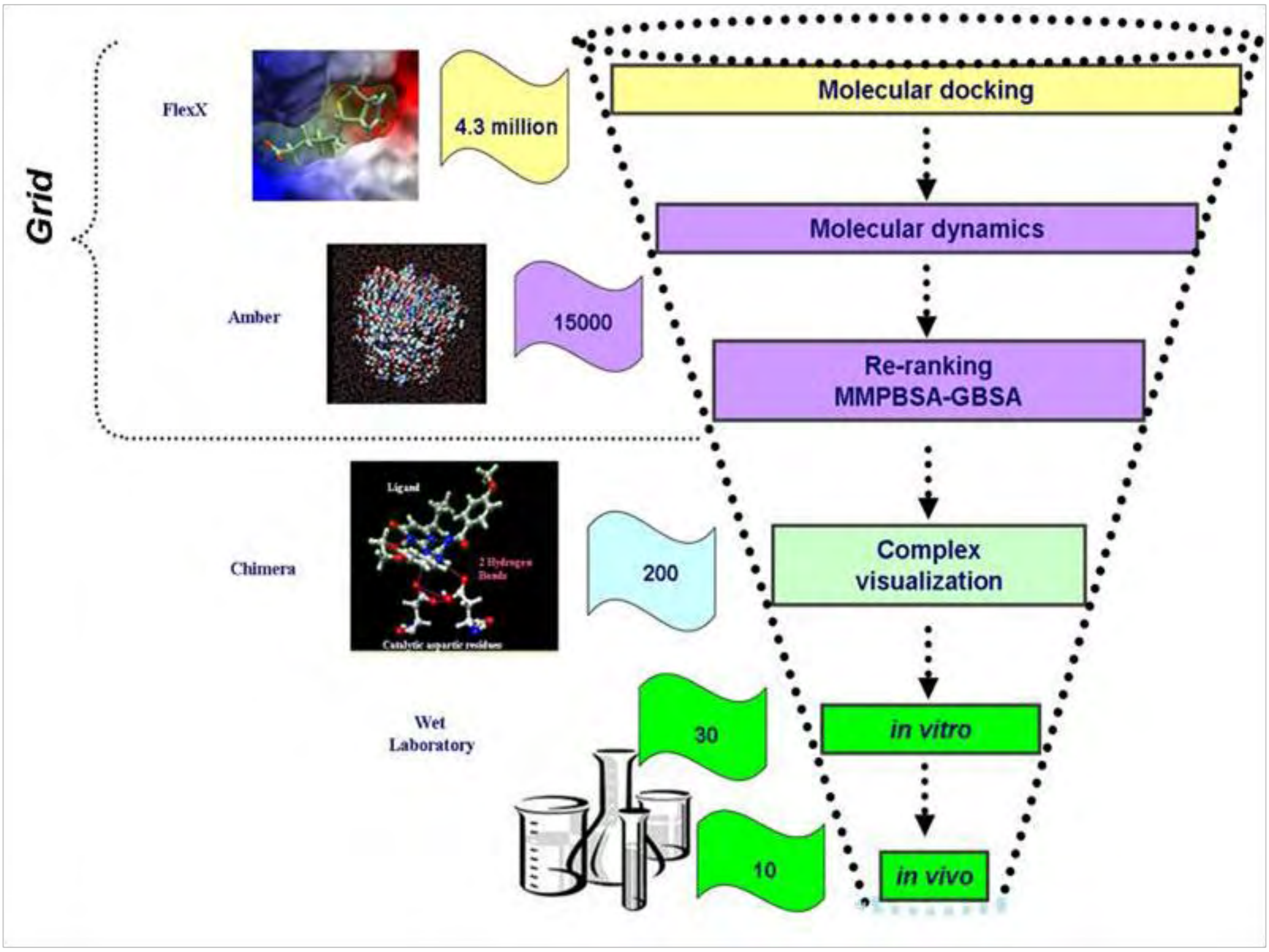


Figure 38: Overall filtering process employed in WISDOM-II project.
Demonstrates overall filtering process employed in WISDOM-II project. The first three steps in the workflow (Docking, Molecular dynamics and rescoring by MMPBSA, MM-GBSA) are performed on Computational Grids and the visualizations by chimera software are performed manually on the local machine of the user.

6.3 Results and Discussion

Docking results

Docking results of PfGST are represented in Table 19. Six amino acids were considered responsible for the activity of the target PfGST: Tyr9, Gln58, Val59, Ser72, Gln71 and Lys15. Chemical compounds interacting with these amino acids were of significance and hence computed. All the ten top scoring compounds displayed in Table19 made interactions to these key amino acids. A binary scoring mode was adopted for the residue reactions in Table 19, column 3: “0” represents false (no interaction with the specified amino acid) and “1” represents true (either a hydrogen bond or a hydrophobic interaction, was made). From Table 19, it is clear that all the top scoring compounds are making interactions with at least one of the key amino acids. These observations are later compared to the standard protein ligand interaction information obtained from ligand plots (www.pdb.org). This particular method not only allowed us to select compounds based on scoring but also based on interaction information (hydrophobic and hydrophilic interactions), which is very significant from the structural point of view for the identification of hits.

Compound	FlexX score	Interaction to Key AA'
ZINC03989574	-50.586	10111
ZINC03989578	-49.698	11101
ZINC04847284	-49.698	11101
ZINC03930012	-48.396	10000
ZINC04522767	-47.633	11100
ZINC05808725	-47.006	01011
ZINC04068384	-46.956	01011
ZINC03948265	-46.286	11100
ZINC02748596	-46.117	11111
ZINC02102883	-46.016	01011

Table 19: Represents top compounds by docking against PfGST with interactions to key amino acids. The Table illustrates the ZINC ids of the top scoring compounds and their docking scores. In the last column, "1" represents a presence of an interaction to key amino acid and "0" represents, no interaction to key amino acids. It was observed that almost all the top scoring compounds made at least one interaction.

6.3.1 Diversity analysis of top scoring compounds for PfGST and PfDHFR

To give wide overview on the results obtained by docking, diversity analysis against the PfGST best scoring 5,000 compounds and PfDHFR best scoring 15,000 compounds was performed by using the MOE software [254]. Fingerprints of all the compounds were created by using FP: BIT MACCS and then used Tanimoto coefficient (TC) for calculating the diversity among the compounds [223]. At similarity cut-off of Tanimoto coefficient 0.7, out of 5,000 compounds of PfGST; 3,394 dissimilar clusters were identified by this method, which further indicates that the best 5,000 compounds diverse and dissimilar. Diversity analysis is performed to demonstrate the best compounds by docking score are sufficiently diverse for the further analysis, and for the identification of novel scaffolds. Pair wise frequency (Y-axis) and Tanimoto coefficient value (X-axis) are plotted, and displayed in Figure 39. The values of mean, median, 1st quartile, 3rd quartile of the histogram are 0.44, 0.43, 0.37, 0.50 respectively. The 1st quartile and 3rd quartile values signify that 25% of the compounds possess TC values of 0.37 and 75% of the compounds possess TC values of 0.5. For PfDHFR diversity analysis is performed against 15,000 top scoring compounds. Pair wise frequency (Y-axis) and Tanimoto coefficient value (TC) (X-axis) are plotted and displayed in Figure 40. The values of mean, median, 1st quartile, 3rd quartile of the histogram are 0.42, 0.40, 0.34, 0.48 respectively. The 1st quartile and 3rd quartile values signify that 25% of the compounds possess TC values of 0.34 and 75% of the compounds possess TC values of 0.48. These observations and figures indicate that the top scoring compounds are diverse and have potential to find novel compounds. The frequency on the Y-axis represents pair wise similarity of each compound against all the compounds in the database (5,000 X 5,000 times for PfGST).

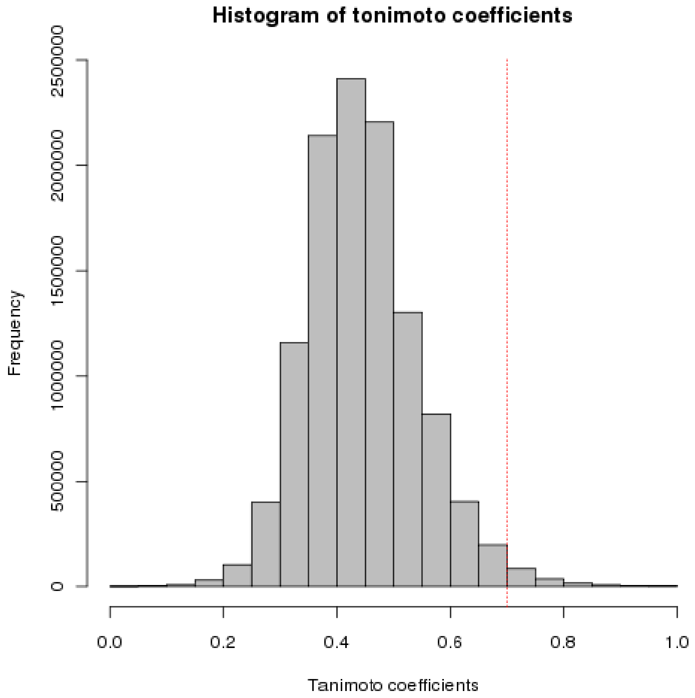


Figure 39: Diversity analysis of the top scoring 5000 compounds against PfGST

.

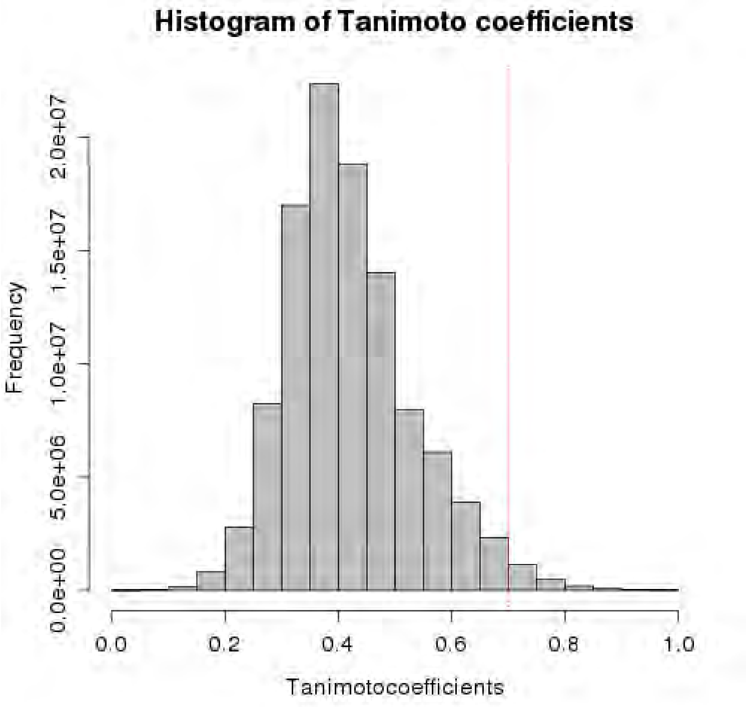


Figure 40: Diversity analysis of the top scoring 15000 compounds against PfDHFR
Demonstrates diversity analysis of the top scoring 15000 compounds against PfDHFR. The red line on the histogram is placed at TC value 0.7 and large bars on the left hand side before the red line indicates, the compound dataset is diverse.

Modeling aspects of final hits against PfGST

To understand the interactions between PfGST and final hits, the ligand plots for each complex (PfGST and the compound) were generated and visualized manually. Protein ligand interactions are studied in three dimensions and for clarity in displaying they are depicted as 2D interaction diagrams. These interactions presented here are generated using the ligand plot module of MOE software. It is evident from Figure 41 that inhibitors are located in the center of the active site, and are stabilized by hydrogen bonding interactions. The hydrogen bonding information along with their distances is listed in Table 20. Figure 41 displays the binding modes of the five best compounds in the active site of the PfGST_a chain. To allow the comparison of binding mode of the compounds and co-crystallized ligand, ligand plot and interactions information is generated for GTX (Cocrystallized ligand of PfGST). It is obvious from Table 20 and Figure 41 that the compounds listed here possess comparable binding poses and patterns. Especially compounds ZINC03533756, ZINC03830430, ZINC03580546, ZINC02305869 generated interaction patterns very similar to the one observed with GTX; making hydrogen bonding to Val59 and Ser72 with backbone as well as with side chains of the amino acids.

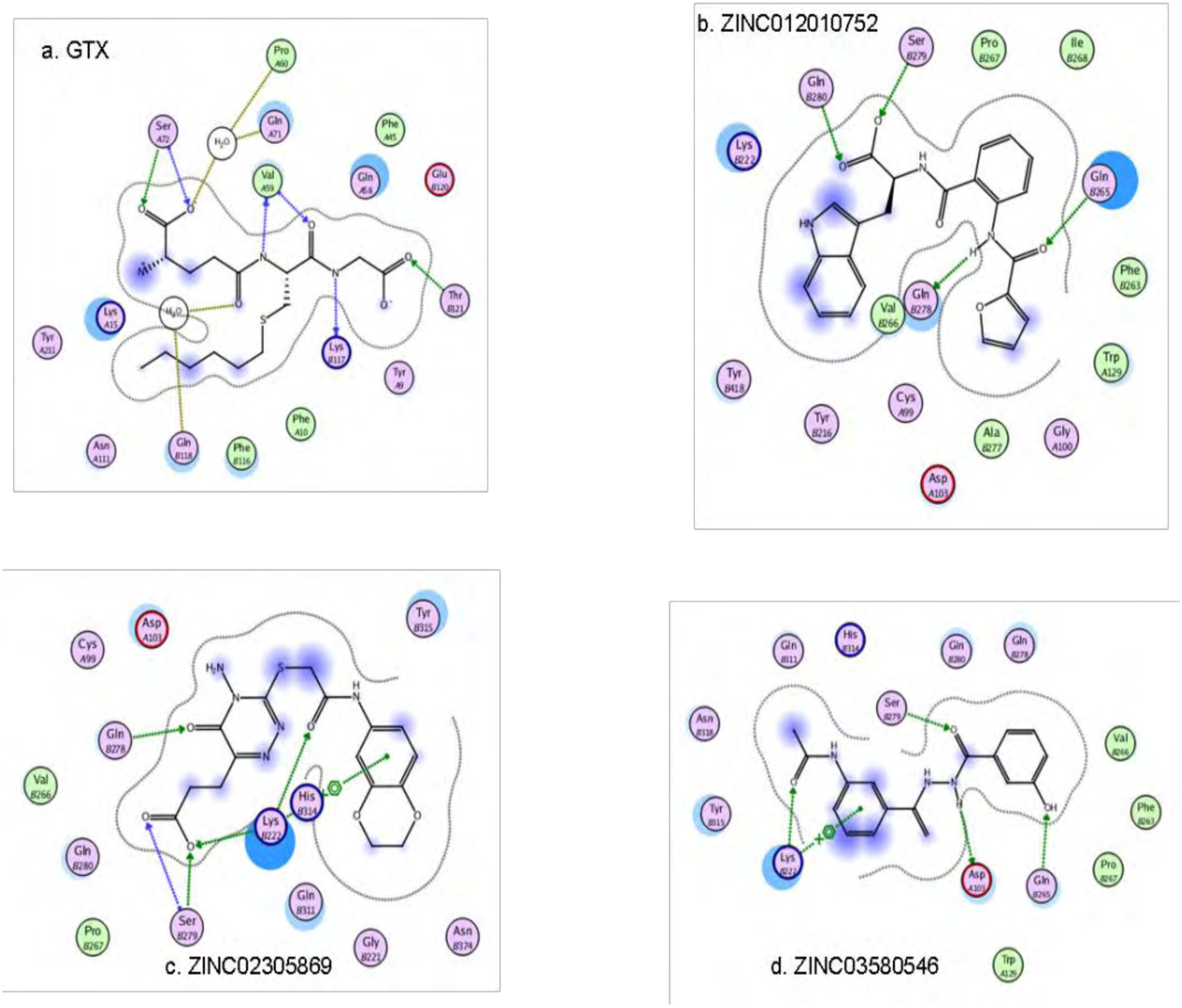


Figure 41: PfGST-compound hydrogen bonding interaction
Interaction informations are displayed for the best compounds which have comparable hydrogen bonding pattern like that of co-crystallized ligand, a.GTX.

Ligand Name	Ligand—Protein	Protein Residue	Type of interaction	Distance Å
GTX	1. N--O & O—N 2. O—OG & O—N 3. N—O	1. Val59 2. SER72 3. LYS117	1. H-don & H-acc 2. H-acc & H-acc 3. H-don	1: 2.85 & 2.81 2: 2.51 & 2.87 3: 2.81
ZINC012010752	1. N—OE & O—NE 2. O—OG 3. O—NE	1. GLN71 2. SER72 3. GLN56	1. H-don & H-acc 2. H-acc 3. H-acc	1: 1.93 & 3.02 2: 2.78 3: 3.02
ZINC01788367	1. O—N & O—OG	1. SER72	1. H-acc & H-acc	1: 3.02 & 2.89
ZINC02305869	1. O—NZ & O—NZ 2. O--NE 3. O—N & O—OG	1. LYS15 2.GLN71 3. SER72	1. H-acc & H-acc 2. H-acc 3. H-acc & H-acc	1. 2.97 & 2.93 2: 2.99 3: 2.89 & 2.83
ZINC02449312	1. O—OG	1. SER72	1. H-acc	1: 2.89
ZINC03533756	1. N--O & O—N 2. O—OG & O—N	1. Val59 2. SER72	1. H-don & H-acc 2. H-acc & H-acc	1: 2.11 & 3.05 2: 3.04 & 2.92
ZINC03580546	1. N—OD 2. O--NE 3. O—OG	1. ASP105 (B) 2. GLN58 3. SER72	1. H-don 2. H-acc 3. H-acc	1: 2.30 2: 3.11 3: 2.99
ZINC03830430	1. O—N 2. O—N	1. Val59 2. SER72	1. H-acc 2. H-acc	1: 2.91 2: 2.92
ZINC05225308	1. O—NZ & O—NZ 2. O--NE 3. O—N & O—OG	1. LYS15 2.GLN71 3. SER72	1. H-acc & H-acc 2. H-acc 3. H-acc & H-acc	1: 2.95 & 3.28 2: 2.92 3: 3.30 & 2.92
ZINC02453649	1. O—NE 2. O—N & O—OG	1. GLN56 2. SER72	1. H-acc 2. H-acc &. H-acc	1: 2.93 2: 2.92 & 2.82

Table 20: PfGST interactions against best compounds are displayed. Especially displays H-bond interactions. In the table, column 2 represents the ligand atom to protein atom interaction, column 3 represents the protein residue against which the compound made the interaction, column 4 represents the type of interaction, column 5 represents the distance at with the H-Bond is formed.

6.4 Summary

The first large-scale docking experiment described in chapter 4 focused on virtual screening against single family of proteins, plasmepsins. However due to complexity of the plasmodium life cycle and drug resistance, multi-target approach is necessary. Hence, in the current chapter, multiple independent targets implicated in malaria, which has different mechanisms of actions are targeted. This chapter describes the collaborative effort taken up to tackle malaria. Various scientific groups all around the world (France, Italy, Venezuela, and South Africa) propose the target proteins screened in this chapter. The research groups that proposed the targets proteins, further shared their knowledge and expertise on the target. The *in silico* screening effort described in this chapter focused on: one new target, glutathione-S-transferase, and two previously well known: dihydrofolate reductase from *Plasmodium falciparum* and *Plasmodium vivax.*

In view of this collaborative effort, it is worth mentioning that, terra bytes of scientific data is shared efficiently and securely on EGEE Grid. This illustrates that, besides the use of the computational Grids for producing large amount of scientific data, Grids form a platform for the convenient global exchange of the chemical data produced.

Gained from the experience of WISDOM-I (Chapter 4), several enhancements to the workflow have been made on both the technical and the modeling side. On the technical side, one major bottleneck in large-scale screening experiments was the handling of large data output of these experiments. During the WISDOM-I project, the large-scale data output was analyzed by scripting languages (perl and shell); this was labor intensive and further significantly delayed the result analysis. As shown, in this chapter, we addressed this problem by parsing the results into a MySQL database. Important information such as the docking score, as well as atom-to-atom interaction between the protein and ligand are stored. The interaction information plays a vital role in selecting the hits, since it takes the compound counterpart, the protein, into consideration as well.

On the modeling side, diversity analysis (using fingerprints and Tanimoto Coefficient) was performed on the best scoring compounds. This revealed that the compounds are quite diverse and sensible for further analysis to find novel compounds. Further, the best scoring compounds were subjected to molecular dynamics simulations and rescoring by MM-PBSA/GBSA. Several promising compounds were identified against PfGST and PfDHFR.

The research group that proposed the target protein will carry on the final development of drug candidates. For example, further development of inhibitors against PfGST and Pf/PvDHFR will be undertaken by University of Pretoria, South Africa and University of Modena, Italy respectively. The drug itself would go into the public domain, for generic manufacturers to produce. This would achieve the goal of getting new medicines to those who need them at the lowest possible price.

Future works aims at two things: an extension of the virtual screening pipeline by additional analysis methods and an even tighter integration of *in silico* prediction of candidate molecules and experimental validation of the compounds.

7 Chapter 7. Conclusions and Outlook

The e-Science paradigm is based on the observation that an increasing number of scientific problem solving approaches require large computational efforts. One of the key features of e-Science is that it supports scientific work through mediating collaboration between individual researchers in virtual organizations and that it enables large-scale experimentation through the sharing of resources. The current thesis, which is a part of WISDOM project, is one example for the successful utilization of this paradigm in the area of computational life sciences. Making use of the world's largest scientific compute infrastructure, the EGEE grid, we have realized a large virtual screening project aiming at the identification of new, potential candidate molecules against multiple targets encoded by *Plasmodium falciparum,* the malaria causing protozoan parasite. Besides the demonstration that a global e-Science production infrastructure such as EGEE enables a new dimension of *in silico* experimentation in the area of computational life sciences, we were aiming at identifying real new candidate inhibitor molecules that could be proposed for further drug development.

Problem of malaria

Malaria together with many other tropical and protozoan diseases is one of the most neglected diseases by the developed countries as well as by the pharmaceutical industries. DNDi identifies new drug discovery and development is the potential gap in the treatment of malaria. As the malaria is affected to people living in poor countries, due to lack of drug research and development facilities in these countries, they continue to suffer; consequently, every day number of people being affected is increasing. Due to very high costs associated with the drug discovery process, as well as due to late stage attrition rates, novel and cost effective strategies are absolutely needed for combating the neglected diseases, especially malaria.

Grid-enabled virtual screening

In silico screening of chemical compounds against a particular target is termed as Virtual Screening (VS). The costs associated to the virtual screening of chemical compounds are significantly reduced when compared to screening of compounds in experimental laboratory. Beside the costs, virtual screening is fast and reliable. However, it is computationally intensive: docking and simulating a single compound within the active site of a given receptor requires about ~25 minutes on a single processor. With the development of combinatorial chemistry technology, millions of different chemical compounds are now available in electronic databases. To screen all these compounds and store the results is a real data challenge. To address this problem computational grid infrastructures are used.

The current thesis describes the drug finding effort by using *in silico* drug discovery techniques on computational Grid.

- Deployment of complex workflows (docking and molecular dynamics) on computational Grid.
- Grid-enabled virtual screening by molecular docking against plasmepsin family of proteins.
- Grid-enabled molecular dynamics simulations for rescoring of docking conformations.
- Large-scale collaborative drug discovery effort for finding hits against multiple targets implicated in malaria (Multi-target approach).

7.1 Discussion of research results

Docking and molecular dynamics were employed for hit and lead finding, as a part of World-wide *In silico* Docking On Malaria (WISDOM) project. WISDOM is one of the first biomedical applications on computational Grids, which used virtual screening by molecular docking and molecular dynamics to find novel drugs against malaria and other neglected diseases.

Virtual screening against plasmepsin family of proteins

As a first attempt, the first WISDOM project (WISDOM-I described in chapter 4), one million chemical compounds obtained from the ChemBridge database, a subset of ZINC, were screened against five targets of the plasmepsin family by using FlexX and AutoDock docking software. 41 million dockings were recorded in 45 days, which is equivalent to 80 CPU years (Grids aspects of WISDOM-I are not a part of this thesis). On the biological side three novel chemotypes against plasmepsin, an aspartic protease were discovered and validated in experimental laboratories. With the new family of potential inhibitors, the guanidino group of compounds, we have established a new class of chemical entities with inhibitory activity against *P. falciparum* plasmepsins. A strong support for their putative activity is that most of the so far known antimalarial drugs likewise contain basic groups and the fact that we identify candidate inhibitors that fall into the already well-established inhibitor classes of thiourea and diphenyl urea analogues speak for the route we have taken. Several potential issues were identified during the large-scale docking experiment, both on the technical side and on modelling side. On the technical side, handling massive docking data (virtual screening results) as flat files was a huge challenge. A customized docking database such as docking database (DDB) from BioSolveIT Gmbh would be an ideal choice for storing and analyzing the results. On the modeling side, due to the robust nature of the docking algorithm and scoring function, significant parameters such as protein flexibility and solvent parameters were ignored. This may be one of the reasons, why several compounds were having huge van der Waals clashes with the receptor atoms.

Combination of molecular docking and molecular dynamics simulations

To overcome the modelling issues, molecular dynamics simulations were performed on the best-docked conformations resulting from WISDOM (Described in Chapter 5). Molecular dynamics addresses electrostatic solvation parameters, protein flexibility and additional degrees of freedom for both protein and ligand. Consequently, it requires more CPU time compared to docking. Hence, molecular dynamics is applied to a restricted number of compounds, usually the best hits coming out of the docking step of the WISDOM project. Again, computational Grids appear very promising to improve the virtual screening process by rescoring using molecular dynamics simulations. Binding free energy calculations were performed utilizing widely used scoring functions: MM-PBSA and MM-GBSA methods of the AMBER9 software. Molecular dynamic simulations were performed initially on 5000 docked conformations of plasmepsin. The compounds were simulated in 7 days which otherwise would have taken 70 days on a single CPU (Described in Chapter 3). The first attempt at using Grids for large-scale virtual screening (combination of molecular docking and molecular dynamics) against plasmepsin ended up in the identification of previously unknown scaffolds, which were confirmed in vitro to be active plasmepsin inhibitors. The combination of docking and molecular dynamics simulations, followed by rescoring using sophisticated scoring functions resulted in the identification of 26 novel sub-micromolar inhibitors. The inhibitors are further clustered into five different scaffolds. While two scaffolds, diphenyl urea and thiourea analogues are already known as plasmepsin inhibitors, albeit the compounds identified here are different from the existing ones, with the new class of potential inhibitors, the guanidino group of compounds, we have established a new class of chemical entities with inhibitory activity against *P. falciparum* plasmepsins.

Collaborative effort to tackle malaria

The success of WISDOM-I led to a succeeding project against malaria: WISDOM-II (Described in Chapter 6). Collaborations were established with research groups from Germany, France, Italy, South Africa, Venezuela and South Korea to select targets and to perform in vitro tests after VS. This time, the focus is to address malaria with a “multi drug therapy”. Therefore, multiple targets from multiple species of Plasmodium (*P. falciparum* and *P. vivax*) were examined. One new target: Glutathione-S-transferase and two previously well-known proteins: dihydrofolate reductase from *P. falciparum* and *P. vivax* were tested. Approximately 4.3 million chemical compounds obtained from the ZINC database were docked against crystal structures of above mentioned targets. To overcome the large-scale data analysis problem, all docking results are stored in a MySQL database and on a central storage element of the Grid, enabling all research groups in the collaboration to access the data. Molecular dynamic simulations were performed on 5000 and 15,000 best-docked conformations of PfGST and PfDHFR, respectively. Twenty thousand compounds were simulated in 18 days which otherwise would have taken 277 days on a single CPU. The modelling results against PfGST and PfDHFR and PvDHFR are quite promising and the research groups, which proposed these target proteins, are further developing the leads into drugs.

This thesis is not only an example where biomedical applications such as molecular docking and molecular dynamics workflows were deployed successfully. It moreover demonstrated how computational Grids could be utilized for producing, storing and sharing terra bytes of scientific data across different research organizations located in different parts of the world with the common goal of finding drugs against malaria.

As described in the chapter 1, the main goals of this thesis are to discover novel small molecules against malaria and to demonstrate the significance of computational Grids in biomedical applications; both the goals have been successfully achieved. On the computational side, molecular docking and molecular dynamics applications were deployed on computational Grids and, on the biological side, novel inhibitors against plasmepsin, an aspartic protease, were identified.

7.2 Outlook

In this thesis, the potential impact of Grid infrastructures for *in silico* drug discovery is demonstrated. The effort described here focused on four malaria biological targets, Pfplasmepsin, PvDHFR, PfDHFR and GST, but at much reduced cost, the same strategy can be applied to produce focused compound libraries for any other malaria targets. Through this thesis, the intention is to draw the attention of the research communities working on these neglected diseases to the opportunity offered by this Grid-enabled virtual screening approach for producing short lists of particularly promising molecules, which can be tested in vitro at a reduced cost.

One major bottleneck in large-scale screening experiments is the handling of large data output of these experiments. As shown chapter 6, this problem is addressed by parsing the results into a MySQL database, the results stored are: the docking scores, as well as atom-to-atom interaction between the protein and ligand. The interaction information plays a vital role in selecting the hits, since it takes the compound counterpart, the protein, into consideration as well.

As an extension of the *in silico* pipeline for virtual screening, data handling and data analysis methods have to be improved significantly. The storage of docking results in the database was just a first step; in the future it is expected to be able to learn from *in silico* experiments by analysing entire series of docking experiments. Techniques supporting the judicious selection of chemical compounds from the large-scale screening data will need to be improved.

New features of drug-like molecules such as their potential toxicity will have to be addressed by an extension of the *in silico* screening through predictive toxicology systems. On the long run, it is likely to extend the *in silico* drug discovery workflow by models for predictive ADME. The rather proprietary nature of the drug discovery process in the pharmaceutical industry resulted in limited availability of models in this field, but initiatives such as the European Innovative Medicine Initiative (IMI) might help to foster broader uptake of computational models for predictive ADME (and toxicity) by altruistic research initiatives, such as WISDOM.

The current thesis may serve as a template for finding hits cost effectively by utilizing the *in silico* methods against multiple targets at the same time. The WISDOM collaboration is also keen to receive requests for docking other malarial targets according to the procedure described in this thesis.

8 Bibliography

[1] P. J. Hotez, D. H. Molyneux, A. Fenwick, J. Kumaresan, S. E. Sachs, J. D. Sachs, and L. Savioli, “Control of neglected tropical diseases.,” *N Engl J Med*, vol. 357, pp. 1018–1027, Sep 2007.

[2] A. R. Renslo and J. H. McKerrow, “Drug discovery and development for neglected parasitic diseases.,” *Nat Chem Biol*, vol. 2, pp. 701–710, Dec 2006.

[3] “Distribution of tropical diseases,” 2005. http://www.pharmabiz.com/article/detnews.asp?articleid=11217&sectionid=48.

[4] P. Trouiller, P. Olliaro, E. Torreele, J. Orbinski, R. Laing, and N. Ford, “Drug development for neglected diseases: a deficient market and a public-health policy failure.,” *Lancet*, vol. 359, pp. 2188–2194, Jun 2002.

[5] P. Chirac and E. Torreele, “Global framework on essential health r&d.,” *Lancet*, vol. 367, pp. 1560–1561, May 2006.

[6] R. Carter and K. N. Mendis, “Evolutionary and historical aspects of the burden of malaria.,” *Clin Microbiol Rev*, vol. 15, pp. 564–594, Oct 2002.

[7] “World malaria report,” 2008. http://www.who.int/malaria/wmr2008/MAL2008-SumKey-EN.pdf.

[8] “Center for diseases control,” 2005. http://www.cdc.gov/malaria/distribution_epi/distribution.htm.

[9] “Malaria distribution,” 2005. http://www.malariasite.com/malaria/WhatIsMalaria.htm.

[10] “World health organization,” 2006. http://www.who.int/tdr/diseases/malaria/default.htm.

[11] “Malaria life cycle,” 2008. www3.niaid.nih.gov/topics/Malaria/lifecycle.htm.

[12] “Parasites and health: Malaria,” May 2004. http://www.dpd.cdc.gov/dpdx/HTML/Malaria.htm.

[13] D. John, W. Petri, and V. Markell, *Markell and Voge's medical parasitology*. St. Louis: Saunders Elsevier, 2006.

[14] A. F. Slater, W. J. Swiggard, B. R. Orton, W. D. Flitter, D. E. Goldberg, A. Cerami, and G. B. Henderson, “An iron-carboxylate bond links the heme units of malaria pigment.,” *Proc Natl Acad Sci U S A*, vol. 88, pp. 325–329, Jan 1991.

[15] S. Pagola, P. W. Stephens, D. S. Bohle, A. D. Kosar, and S. K. Madsen, “The structure of malaria pigment beta-haematin.,” *Nature*, vol. 404, pp. 307–310, Mar 2000.

[16] S. Pukrittayakamee, A. Chantra, S. Vanijanonta, R. Clemens, S. Looareesuwan, and N. J. White, “Therapeutic responses to quinine and clindamycin in multidrug-resistant falciparum malaria.,” *Antimicrob Agents Chemother*, vol. 44, pp. 2395–2398, Sep 2000.

[17] S. Pukrittayakamee, A. Chantra, J. A. Simpson, S. Vanijanonta, R. Clemens, S. Looareesuwan, and N. J. White, “Therapeutic responses to different antimalarial drugs in vivax malaria.,” *Antimicrob Agents Chemother*, vol. 44, pp. 1680–1685, Jun 2000.

[18] S. Pukrittayakamee, K. Chotivanich, A. Chantra, R. Clemens, S. Looareesuwan, and N. J. White, “Activities of artesunate and primaquine against asexual- and sexual-stage parasites in falciparum malaria.,” *Antimicrob Agents Chemother*, vol. 48, pp. 1329–1334, Apr 2004.

[19] N. J. White, “Drug resistance in malaria.,” *Br Med Bull*, vol. 54, no. 3, pp. 703–715, 1998.

[20] N. J. White, F. Nosten, S. Looareesuwan, W. M. Watkins, K. Marsh, R. W. Snow, G. Kokwaro, J. Ouma, T. T. Hien, M. E. Molyneux, T. E. Taylor, C. I. Newbold, T. K. Ruebush, M. Danis, B. M. Greenwood, R. M. Anderson, and P. Olliaro, “Averting a malaria disaster.,” *Lancet*, vol. 353, pp. 1965–1967, Jun 1999.

[21] R. N. Price, F. Nosten, C. Luxemburger, F. O. ter Kuile, L. Paiphun, T. Chongsuphajaisiddhi, and N. J. White, “Effects of artemisinin derivatives on malaria transmissibility.,” *Lancet*, vol. 347, pp. 1654–1658, Jun 1996.

[22] L. J. Bruce-Chwatt, “Malaria, the growing medical and health problem.,” *Drugs Exp Clin Res*, vol. 11, no. 12, pp. 899–909, 1985.

[23] L. J. Bruce-Chwatt, “Malaria and its control: present situation and future prospects.,” *Annu Rev Public Health*, vol. 8, pp. 75–110, 1987.

[24] N. J. White, “Antimalarial drug resistance.,” *J Clin Invest*, vol. 113, pp. 1084–1092, Apr 2004.

[25] D. J. Walker, J. L. Pitsch, M. M. Peng, B. L. Robinson, W. Peters, J. Bhisutthibhan, and S. R. Meshnick, “Mechanisms of artemisinin resistance in the rodent malaria pathogen plasmodium yoelii.,” *Antimicrob Agents Chemother*, vol. 44, pp. 344–347, Feb 2000.

[26] S. Thaithong, “Clones of different sensitivities in drug-resistant isolates of plasmodium falciparum.,” *Bull World Health Organ*, vol. 61, no. 4, pp. 709–712, 1983.

[27] P. G. Kremsner and S. Krishna, “Antimalarial combinations.,” *Lancet*, vol. 364, no. 9430, pp. 285–294, 2004.

[28] K. Buse and G. Walt, “Global public-private partnerships: Part ii–what are the health issues for global governance?,” *Bull World Health Organ*, vol. 78, no. 5, pp. 699–709, 2000.

[29] K. Buse and G. Walt, “Global public-private partnerships: Part i–a new development in health?,” *Bull World Health Organ*, vol. 78, no. 4, pp. 549–561, 2000.

[30] “Walter reed army insitute of research,” 2004. http://wrair-www.army.mil/.

[31] “Research and training in tropical diseases.” http://www.who.int/tdr/diseases/malaria/default.htm.

[32] “Drugs for neglected diseases initiative,” 2006. http://www.dndi.org/.

[33] “Malaria vaccine,” 2005. http://www.malariavaccine.org.

[34] “Medicines for malaria venture,” 2005. http://www.mmv.org/rubrique.php3?id_rubrique=15.

[35] “Roll back malaria,” 2005. http://www.rbm.who.int/cgi-bin/rbm/rbmportal/custom/rbm/home.do.

[36] “Wellcome trust,” 2004. http://www.wellcome.ac.uk/.

[37] “St. jude medical center,” 2004. http://www.stjudemedicalcenter.org/.

[38] D. A. Fidock, P. J. Rosenthal, S. L. Croft, R. Brun, and S. Nwaka, “Antimalarial drug discovery: efficacy models for compound screening.,” *Nat Rev Drug Discov*, vol. 3, pp. 509–520, Jun 2004.

[39] R. W. Snow, C. A. Guerra, J. J. Mutheu, and S. I. Hay, “International funding for malaria control in relation to populations at risk of stable plasmodium falciparum transmission.,” *PLoS Med*, vol. 5, p. e142, Jul 2008.

[40] K. H. Bleicher, H.-J. Böhm, K. Müller, and A. I. Alanine, “Hit and lead generation: beyond high-throughput screening.,” *Nat Rev Drug Discov*, vol. 2, pp. 369–378, May 2003.

[41] J. Bajorath, “Integration of virtual and high-throughput screening.,” *Nat Rev Drug Discov*, vol. 1, pp. 882–894, Nov 2002.

[42] H.-J. Böhm, G. Schneider, H. Kubinyi, R. Mannhold, and H. Timmerman, *Virtual Screening For Bioactive Molecules (Methods and Principles in Medicinal Chemistry)*. Wiley-VCH, 2003.

[43] T. Lengauer, *Bioinformatics from genome To Drugs*. Wiley-VCH, 2001.

[44] H. D. Hoeltje, W. Sippl, D. Rognan, and G. Folkers, *Molecular Modeling, Basics Principles And Applications*. Wiley-VCH, 2003.

[45] B. K. Shoichet, S. L. McGovern, B. Wei, and J. J. Irwin, “Lead discovery using molecular docking.,” *Curr Opin Chem Biol*, vol. 6, pp. 439–446, Aug 2002.

[46] M. Y. Mizutani and A. Itai, “Efficient method for high-throughput virtual screening based on flexible docking: discovery of novel acetylcholinesterase inhibitors.,” *J Med Chem*, vol. 47, pp. 4818–4828, Sep 2004.

[47] R. Li, X. Chen, B. Gong, P. M. Selzer, Z. Li, E. Davidson, G. Kurzban, R. E. Miller, E. O. Nuzum, J. H. McKerrow, R. J. Fletterick, S. A. Gillmor, C. S. Craik, I. D. Kuntz, F. E. Cohen, and G. L. Kenyon, “Structure-based design of parasitic protease inhibitors.,” *Bioorg Med Chem*, vol. 4, pp. 1421–1427, Sep 1996.

[48] P. D. Lyne, “Structure-based virtual screening: an overview.,” *Drug Discov Today*, vol. 7, pp. 1047–1055, Oct 2002.

[49] H. Alonso, A. A. Bliznyuk, and J. E. Gready, “Combining docking and molecular dynamic simulations in drug design.,” *Med Res Rev*, vol. 26, pp. 531–568, Sep 2006.

[50] C. Foster, I.; Kesselman, *Computational Grids. In The Grid: Blueprint for a New Computing Infrastructure*. Morgan Kaufmann Publishers: San Fransisco, CA, U.S.A, 1999.

[51] “The smallpox reasearch grid,” 2006. http://grid.org/projects/smallpox/ index.htm (accessed May 9, 2006).

[52] “The anthrax research project.,” 2006. http://grid.org/projects/anthrax/ index.htm.

[53] “United devices cancer research project.,” 2006. http://grid.org/projects/cancer/ index.htm.

[54] W. G. Richards, “Virtual screening using grid computing: the screensaver project.,” *Nat Rev Drug Discov*, vol. 1, pp. 551–555, Jul 2002.

[55] R. Shankar, X. Frapaise, and B. Brown, “Lean drug development in r&d,” *Drug Discov. Dev.*, p. 57–60, 2006.

[56] I. M. Kapetanovic, “Computer-aided drug discovery and development (caddd): in silico-chemico-biological approach.,” *Chem Biol Interact*, vol. 171, pp. 165–176, Jan 2008.

[57] "Pricewaterhousecoopers, pricewaterhousecoopers pharma 2005: An industrial revolution in r&d.," 2005. http://www.pwc.com/gx/eng/about/ind/pharma/industrial_revolution.pdf.

[58] B. Waszkowycz, "Structure-based approaches to drug design and virtual screening.," *Curr Opin Drug Discov Devel*, vol. 5, pp. 407–413, May 2002.

[59] H. Kubinyi, "Chance favors the prepared mind–from serendipity to rational drug design.," *J Recept Signal Transduct Res*, vol. 19, no. 1-4, pp. 15–39, 1999.

[60] J. C. Lizhe Wang, Wei Jie, *Grid Computing: Infrastructure, Service, and Applications*. CRC; 1 edition, 2008.

[61] J. T. Koh, "Making virtual screening a reality.," *Proc Natl Acad Sci U S A*, vol. 100, pp. 6902–6903, Jun 2003.

[62] H. Köppen, "Virtual screening - what does it give us?," *Curr Opin Drug Discov Devel*, vol. 12, pp. 397–407, May 2009.

[63] S. Subramaniam, M. Mehrotra, and D. Gupta, "Virtual high throughput screening (vhts) - a perspective.," *Bioinformation*, vol. 3, no. 1, pp. 14–17, 2008.

[64] U. Rester, "From virtuality to reality - virtual screening in lead discovery and lead optimization: a medicinal chemistry perspective.," *Curr Opin Drug Discov Devel*, vol. 11, pp. 559–568, Jul 2008.

[65] C. McInnes, "Virtual screening strategies in drug discovery.," *Curr Opin Chem Biol*, vol. 11, pp. 494–502, Oct 2007.

[66] A. S. Reddy, S. P. Pati, P. P. Kumar, H. N. Pradeep, and G. N. Sastry, "Virtual screening in drug discovery – a computational perspective.," *Curr Protein Pept Sci*, vol. 8, pp. 329–351, Aug 2007.

[67] A. N. Jain, "Virtual screening in lead discovery and optimization.," *Curr Opin Drug Discov Devel*, vol. 7, pp. 396–403, Jul 2004.

[68] T. Langer and R. D. Hoffmann, "Virtual screening: an effective tool for lead structure discovery?," *Curr Pharm Des*, vol. 7, pp. 509–527, May 2001.

[69] G. Klebe, "Virtual ligand screening: strategies, perspectives and limitations.," *Drug Discov Today*, vol. 11, pp. 580–594, Jul 2006.

[70] A. R. Leach and V. J. Gillet, *An Introduction to Chemoinformatics*. Kluwer Academic publishers, 2003.

[71] R. Casey, "Bioinformatics in structure-based drug discovery," March 2006. http://www.b-eye-network.com/view/2593.

[72] B. O. Villoutreix, N. Renault, D. Lagorce, O. Sperandio, M. Montes, and M. A. Miteva, "Free resources to assist structure-based virtual ligand screening experiments.," *Curr Protein Pept Sci*, vol. 8, pp. 381–411, Aug 2007.

[73] P. Bamborough and F. E. Cohen, "Modeling protein–ligand complexes," *Current Opinion in Structural Biology*, vol. 6, no. 2, pp. 236 – 241, 1996.

[74] R. Pearlman, *3D Molecular Structures: Generation and Use in 3D Searching, In 3D QSAR in Drug Design: Theory, Methods, and Applications*. ESCOM Science Publishers: Leiden, 1993.

[75] R. D. Cramer, D. E. Patterson, and J. D. Bunce, “Comparative molecular field analysis (comfa). 1. effect of shape on binding of steroids to carrier proteins,” *Journal of the American Chemical Society*, vol. 110, no. 18, pp. 5959–5967, 1988.

[76] D. B. Kitchen, H. Decornez, J. R. Furr, and J. Bajorath, “Docking and scoring in virtual screening for drug discovery: methods and applications.,” *Nat Rev Drug Discov*, vol. 3, pp. 935–949, Nov 2004.

[77] H. Gohlke and G. Klebe, “Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors.,” *Angew Chem Int Ed Engl*, vol. 41, pp. 2644–2676, Aug 2002.

[78] M. Rarey, B. Kramer, T. Lengauer, and G. Klebe, “A fast flexible docking method using an incremental construction algorithm.,” *J Mol Biol*, vol. 261, pp. 470–489, Aug 1996.

[79] I. Kuntz, J. M. Blaney, S. Oatley, R. Langridge, and T. Ferrin, “A geometric approach to macromolecule-ligand interactions.,” *J. Mol. Biol.*, vol. 161, pp. 269–288, 1982.

[80] R. L. DesJarlais, R. P. Sheridan, G. L. Seibel, J. S. Dixon, I. D. Kuntz, and R. Venkataraghavan, “Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure.,” *J Med Chem*, vol. 31, pp. 722–729, Apr 1988.

[81] I. D. Kuntz, “Structure-based strategies for drug design and discovery.,” *Science*, vol. 257, pp. 1078–1082, Aug 1992.

[82] I. Kuntz, E. Meng, and B. Shoichet., “Structure-based molecular design,” *Acc. Chem. Res.*, vol. 27, pp. 117–123, 1994.

[83] C. M. Oshiro and I. D. Kuntz, “Characterization of receptors with a new negative image: use in molecular docking and lead optimization.,” *Proteins*, vol. 30, pp. 321–336, Feb 1998.

[84] “Flexx 2.0,” 2005. http://www.biosolveit.de/.

[85] A. N. Jain, “Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine.,” *J Med Chem*, vol. 46, pp. 499–511, Feb 2003.

[86] W. Welch, J. Ruppert, and A. N. Jain, “Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites.,” *Chem Biol*, vol. 3, pp. 449–462, Jun 1996.

[87] M. Miteva, “Hierarchical structure-based virtual screening for drug design,” *BIOTECHNOL. & BIOTECHNOL*, pp. 634–638, 2008.

[88] K. P. Clark and Ajay, “Flexible ligand docking without parameter adjustment across four ligand–receptor complexes,” *J Comput Chem*, vol. 16, p. 1210–1226, 1995.

[89] G. Jones, P. Willett, and R. C. Glen, “Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation.,” *J Mol Biol*, vol. 245, p. 43–53, 1995.

[90] C. M. Oshiro, I. D. Kuntz, and J. S. Dixon, “Flexible ligand docking using a genetic algorithm.,” *J Comput Aided Mol Des*, vol. 9, pp. 113–130, Apr 1995.

[91] C. M. Venkatachalam, X. Jiang, T. Oldfield, and M. Waldman, “Ligandfit: a novel method for the shape-directed rapid docking of ligands to protein active sites.,” *J Mol Graph Model*, vol. 21, pp. 289–307, Jan 2003.

[92] M. L. Verdonk, J. C. Cole, M. J. Hartshorn, C. W. Murray, and R. D. Taylor, “Improved protein-ligand docking using gold.,” *Proteins*, vol. 52, pp. 609–623, Sep 2003.

[93] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, and A. J. Olson, “Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function,” *Journal of Computational Chemistry*, vol. 19, pp. 1639–1662, January 1999.

[94] J. S. Taylor and R. M. Burnett, “Darwin: a program for docking flexible molecules.,” *Proteins*, vol. 41, pp. 173–191, Nov 2000.

[95] G. Wu and M. Vieth, “Sdocker: a method utilizing existing x-ray structures to improve docking accuracy.,” *J Med Chem*, vol. 47, pp. 3142–3148, Jun 2004.

[96] B. Kuhn, P. Gerber, T. Schulz-Gasch, and M. Stahl, “Validation and use of the mm-pbsa approach for drug discovery.,” *J Med Chem*, vol. 48, pp. 4040–4048, Jun 2005.

[97] M. Liu and S. Wang, “Mcdock: a monte carlo simulation approach to the molecular docking problem.,” *J Comput Aided Mol Des*, vol. 13, pp. 435–451, Sep 1999.

[98] N. Moitessier, P. Englebienne, D. Lee, J. Lawandi, and C. R. Corbeil, “Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go.,” *Br J Pharmacol*, vol. 153 Suppl 1, pp. S7–26, Mar 2008.

[99] H. Gohlke, M. Hendlich, and G. Klebe, “Knowledge-based scoring function to predict protein-ligand interactions.,” *J Mol Biol*, vol. 295, pp. 337–356, Jan 2000.

[100] I. Muegge and Y. C. Martin, “A general and fast scoring function for protein-ligand interactions: a simplified potential approach.,” *J Med Chem*, vol. 42, pp. 791–804, Mar 1999.

[101] H. J. Böhm, “The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure,” *Journal of Computer-Aided Molecular Design,*, vol. 8, pp. 243–256, 1994.

[102] C. A. Reynolds, P. M. King, and W. G. Richards, “Free energy calculations in molecular biophysics,” *Molecular Physics: An International Journal at the Interface Between Chemistry and Physics*, vol. 76, no. 2, pp. 251–275, 1992.

[103] L. Zhang, E. Gallicchio, R. Friesner, and R. Levy., “Solvent models for protein–ligand binding: Comparison of implicit solvent poisson and surface generalized born models with explicit solvent simulations.,” *J. Comput. Chem.*, vol. 22, p. 591–607., 2001.

[104] P. Kirkpatrick and C. Ellis., “Chemical space,” *Nature Rev Drug Discov*, vol. 432, pp. 823–865, 2004.

[105] M. Kontoyianni, L. M. McClellan, and G. S. Sokol, “Evaluation of docking performance: comparative data on docking algorithms.,” *J Med Chem*, vol. 47, pp. 558–565, Jan 2004.

[106] M. Feher, “Consensus scoring for protein-ligand interactions.,” *Drug Discov Today*, vol. 11, pp. 421–428, May 2006.

[107] C. Bissantz, G. Folkers, and D. Rognan, “Protein-based virtual screening of chemical databases. 1. evaluation of different docking/scoring combinations.,” *J Med Chem*, vol. 43, pp. 4759–4767, Dec 2000.

[108] S. Betzi, K. Suhre, B. Chétrit, F. Guerlesquin, and X. Morelli, “Gfscore: a general nonlinear consensus scoring function for high-throughput docking.,” *J Chem Inf Model*, vol. 46, no. 4, pp. 1704–1712, 2006.

[109] A. Wolf, M. Zimmermann, and M. Hofmann-Apitius, “Alternative to consensus scoring–a new approach toward the qualitative combination of docking algorithms.,” *J Chem Inf Model*, vol. 47, no. 3, pp. 1036–1044, 2007.

[110] R. D. Clark, A. Strizhev, J. M. Leonard, J. F. Blake, and J. B. Matthew, “Consensus scoring for ligand/protein interactions.,” *J Mol Graph Model*, vol. 20, pp. 281–295, Jan 2002.

[111] P. S. Charifson, J. J. Corkery, M. A. Murcko, and W. P. Walters, “Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins.,” *J Med Chem*, vol. 42, pp. 5100–5109, Dec 1999.

[112] A. Leach, *Molecular Modeling: Principles and Applications*. Addison Weseley Longmann Limited,, 1996.

[113] J. Ponder and D. Case, “Force fields for protein simulations.,” *Adv. Prot. Chem.*, vol. 66, pp. 27–85, 2003.

[114] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus., “Charmm—a program for macromolecular energy, minimization, and dynamics calculations.,” *J Comput Chem*, vol. 4, p. 187–217, 1983.

[115] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. Berendsen, “Gromacs: fast, flexible, and free.,” *J Comput Chem*, vol. 26, pp. 1701–1718, December 2005.

[116] T. Hansson, C. Oostenbrink, and W. van Gunsteren, “Molecular dynamics simulations.,” *Curr Opin Struct Biol*, vol. 12, pp. 190–196, Apr 2002.

[117] M. Karplus, “Molecular dynamics simulations of biomolecules.,” *Acc Chem Res*, vol. 35, pp. 321–323, Jun 2002.

[118] K. Gibson and H. Scheraga, “Revised algorithms for the build-up procedure for predicting protein conformations by energy minimization,” *J. Comput. Chem.*, vol. 8, pp. 826–834, 1987.

[119] C. B. Anfinsen, “Principles that govern the folding of protein chains.,” *Science*, vol. 181, pp. 223–230, Jul 1973.

[120] C. B. Anfinsen and H. A. Scheraga, “Experimental and theoretical aspects of protein folding.,” *Adv Protein Chem*, vol. 29, pp. 205–300, 1975.

[121] A. R. Leach, “Ligand docking to proteins with discrete side-chain flexibility.,” *J Mol Biol*, vol. 235, pp. 345–356, Jan 1994.

[122] B. Kuhn and P. A. Kollman, “Binding of a diverse set of ligands to avidin and streptavidin: an accurate quantitative prediction of their relative affinities by a combination of molecular mechanics and continuum solvent models.,” *J Med Chem*, vol. 43, pp. 3786–3791, Oct 2000.

[123] J. Wang, P. Morin, W. Wang, and P. A. Kollman, “Use of mm-pbsa in reproducing the binding free energies to hiv-1 rt of tibo derivatives and predicting the binding mode to hiv-1 rt of efavirenz by docking and mm-pbsa.,” *J Am Chem Soc*, vol. 123, pp. 5221–5230, Jun 2001.

[124] B. O. Brandsdal, F. Osterberg, M. Almlöf, I. Feierberg, V. B. Luzhkov, and J. Aqvist, “Free energy calculations and ligand binding.,” *Adv Protein Chem*, vol. 66, pp. 123–158, 2003.

[125] T. P. Straatsma and J. A. McCammon., “Computational alchemy.,” *Annu. Rev. Phys. Chem.,*, vol. 43, pp. 407–435, 1992.

[126] D. L. Beveridge and F. M. DiCapua, “Free energy via molecular simulation: applications to chemical and biomolecular systems.,” *Annu Rev Biophys Biophys Chem*, vol. 18, pp. 431–492, 1989.

[127] J. M. J. Swanson, R. H. Henchman, and J. A. McCammon, “Revisiting free energy calculations: a theoretical connection to mm/pbsa and direct calculation of the association free energy.,” *Biophys J*, vol. 86, pp. 67–74, Jan 2004.

[128] J. Aqvist, C. Medina, and J. E. Samuelsson, “A new method for predicting binding affinity in computer-aided drug design.,” *Protein Eng*, vol. 7, pp. 385–391, Mar 1994.

[129] T. Hansson and J. Aqvist, “Estimation of binding free energies for hiv proteinase inhibitors by molecular dynamics simulations.,” *Protein Eng*, vol. 8, pp. 1137–1144, Nov 1995.

[130] J. Aqvist, “Calculation of absolute binding free energies for charged ligands and effects of long-range electrostatic interactions,” *J Comput Chem*, vol. 17, p. 1587–1597, 1996.

[131] T. Hansson, J. Marelius, and J. Aqvist, “Ligand binding affinity prediction by linear interaction energy methods.,” *J Comput Aided Mol Des*, vol. 12, pp. 27–35, Jan 1998.

[132] H. Park, M. S. Yeom, and S. Lee, “Loop flexibility and solvent dynamics as determinants for the selective inhibition of cyclin-dependent kinase 4: comparative molecular dynamics simulation studies of cdk2 and cdk4.,” *Chembiochem*, vol. 5, pp. 1662–1672, Dec 2004.

[133] A. Cavalli, G. Bottegoni, C. Raco, M. D. Vivo, and M. Recanatini, “A computational study of the binding of propidium to the peripheral anionic site of human acetylcholinesterase.,” *J Med Chem*, vol. 47, pp. 3991–3999, Jul 2004.

[134] G. Rastelli, A. M. Ferrari, L. Costantino, and M. C. Gamberini, “Discovery of new inhibitors of aldose reductase from molecular docking and database screening.,” *Bioorg Med Chem*, vol. 10, pp. 1437–1450, May 2002.

[135] S. Hammer, I. Spika, W. Sippl, G. Jessen, B. Kleuser, H. Holtje, and M. Schafer-Korting., “Glucocorticoid receptor interactions with glucocorticoids: Evaluation by molecular modeling and functional analysis of glucocorticoid receptor mutants.,” *Steroids*, vol. 68, p. 329–339, 2003.

[136] C. E. Cannizzaro, J. A. Ashley, K. D. Janda, and K. N. Houk, “Experimental determination of the absolute enantioselectivity of an antibody-catalyzed diels-alder reaction and theoretical explorations of the origins of stereoselectivity.,” *J Am Chem Soc*, vol. 125, pp. 2489–2506, Mar 2003.

[137] R. García-Nieto, C. Pérez, and F. Gago, “Automated docking and molecular dynamics simulations of nimesulide in the cyclooxygenase active site of human prostaglandin-endoperoxide synthase-2 (cox-2).,” *J Comput Aided Mol Des*, vol. 14, pp. 147–160, Feb 2000.

[138] I. Foster, C. Kesselman, and S. Tuecke, “The anatomy of the grid: Enabling scalable virtual organizations,,” *International Journal of High Performance Computing Applications,*, vol. 15, pp. 200–222, 2001.

[139] I. Foster, “What is the grid? - a three point checklist,” *GRIDtoday*, vol. 1, July 2002.

[140] M. Chetty and R. Buyya, “Weaving computational grids: how analogous are they with electrical grids?,” *Computing in Science & Engineering*, vol. 4, pp. 61–71, July–Aug. 2002.

[141] V. Breton, R. Medina, and J. Montagnat, “Datagrid, prototype of a biomedical grid.,” *Methods Inf Med*, vol. 42, no. 2, pp. 143–147, 2003.

[142] A. Wolf, M. Shahid, V. Kasam, W. Ziegler, and M. Hofmann-apitius, “In silico drug discovery approaches on grid computing infrastructures,” *Current Clinical Pharmacology*, 2009.

[143] J. Nicolas, *In silico drug discovery services in computing grid environments against neglected and emerging infectious diseases*. PhD thesis, UNIVERSITE BLAISE PASCAL, 2006.

[144] “The eugridpma - coordinating grid authentication in e-science,” 2006. http://www.eugridpma.org/.

[145] N. Jacq, J. Salzemann, F. Jacq, Y. Legré, E. Medernach, J. Montagnat, A. Maaß, M. Reichstadt, H. Schwichtenberg, M. Sridhar, V. Kasam, M. Zimmermann, M. Hofmann, and V. Breton, “Grid-enabled virtual screening against malaria.,” *J. Grid Comput.*, vol. 6(1), pp. 29–43, 2008.

[146] N. Jacq, V. Breton, H-Y.Chen, L.-Y. Ho, M. Hofmann, H.-C. Lee, Y. Legré, S. C. Lin, A. Maaß, E. Medernach, I. Merelli, L. Milanesi, G. Rastelli, M. Reichstadt, J. Salzemann, H. Schwichtenberg, M. Sridhar, V. Kasam, Y.-T. Wu, and M. Zimmermann., “Virtual screening on large scale grids.,” *Parallel Comput*, vol. 33, p. 289–301., 2007.

[147] I. Foster, C. Kesselman, G. Tsudik, and S. Tuecke, “A security architecture for computational grids.,” *In ACM Conference on Computers and Security*, 1998.

[148] “Fightaids@home,” 2008. http://fightaidsathome.scripps.edu/index.html.

[149] “Seti@home,” 2008. http://setiathome.berkeley.edu/.

[150] “Enabling science for e-science,” 2006. http://www.eu-egee.org/.

[151] D. Anderson and J. Kubiatowicz, “The worldwide computer,” *Sci.Am.*, vol. 3, pp. 40–47, 2002.

[152] M. Mutka and M. Livny, “The available capacity of a privately owned workstation environment.,” *Performance Evaluation*, vol. 12, p. 269–284, 2002.

[153] K. Ryu and J. Hollingsworth, “Exploiting fine grained idle periods in networks of workstations,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 11, pp. 683–698, 2000.

[154] L. Loewe, “Global computing for bioinformatics,” *Briefings in Bioinformatics*, vol. 3, pp. 377–388, 2002.

[155] A. Oram, ed., *Peer-to-Peer: Harnessing the Power of Disruptive Technologies*. O’Reilly Press, 2001.

[156] D. Anderson and G. Fedak, “The computational and storage potential of volunteer computing,,” in *IEEE/ACM International Symposium on Cluster Computing and the Grid*, 2006.

[157] Y. Shavitt and E. Shir, “Dimes: Let the internet measure itself,” *Computer Communication*, vol. 5, pp. 71–74, 2005.

[158] “Grid computing with boinc,” 2006. http://boinc.berkeley.edu/trac/wiki/DesktopGrid.

[159] R. Buyya, “Economic models for management of resources in peer-to-peer and grid computing,,” in *Proc. SPIE Int’l Conf. on Commercial Applications for High-Performance Computing*, 2001.

[160] M. Baker, “The grid: International efforts in global computing,” in *Proc. Int’l Conf. Advances in Infrastructure for Electronic Business, Science, and Education on the Internet*, 2000.

[161] M. Hewett, D. E. Oliver, D. L. Rubin, K. L. Easton, J. M. Stuart, R. B. Altman, and T. E. Klein, “Pharmgkb: the pharmacogenetics knowledge base.,” *Nucleic Acids Res*, vol. 30, pp. 163–165, Jan 2002.

[162] “E-science grid facility for europe and latin america,” 2006. www.eu-eela.org.

[163] “Euchinagrid,” 2006. www.euchinagrid.org.

[164] “Mygrid,” 2008. http://www.mygrid.org.uk/.

[165] H. Rauwerda, M. Roos, B. O. Hertzberger, and T. M. Breit, “The promise of a virtual lab in drug discovery.,” *Drug Discov Today*, vol. 11, pp. 228–236, Mar 2006.

[166] W. E. Johnston, D. Gannon, and B. Nitzberg, “Grids as production computing environments: the engineering aspects of nasa's information power grid,” in *Proc. Eighth International Symposium on High Performance Distributed Computing*, pp. 197–204, 3–6 Aug. 1999.

[167] B. R, “The world-wide grid,” 2008. http://www.buyya.com/ecogrid/wwg/.

[168] “Nsf tera-grid,” 2008. http://www.teraGrid.org/.

[169] W. Sanchez, B. Gilman, M. Kher, S. Lagou, and P. Covitz., “cagrid white paper national cancer institute center for bioinformatics,” tech. rep., National Cancer Institute, 2004.

[170] I. Foster, C. Kesselman, J. M. Nick, and S. Tuecke., “Grid services for distributed system integration,” *Computer*, vol. 35,, pp. 37–46, 2002.

[171] I. Foster, C. Kesselman, J. Nick, and S. Tuecke., “Physiology of the grid..” Available from http://www.globus.org/research/papers/ogsa.pdf.

[172] “Web service resource framework,” 2006. http://www.globus.org/wsrf/.

[173] “Web services description language,” 2007. http://www.w3.org/TR/wsdl.

[174] A. Y. Z. El-Ghazali Talbi, ed., *Grid computing for Bioinformatics and computational biology*. Wiley-Interscience, 2007.

[175] S. Krishnan and K. Bhatia, “Soas for scientific applications: Experiences and challenges,” in *Proc. IEEE International Conference on e-Science and Grid Computing*, pp. 160–169, 10–13 Dec. 2007.

[176] C. Combet, C. Blanchet, C. Geourjon, and G. Deléage, “Nps@: network protein sequence analysis.,” *Trends Biochem Sci*, vol. 25, pp. 147–150, Mar 2000.

[177] “Institut de biologie et chimie des proteines. nps@: Welcome to network protein sequence @nalysis at ibcp, france,” 2008 Sep 2. http://gpsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_server.html.

[178] “Dockingserver,” 2008. http://www.dockingserver.com/.

[179] “Charmming web interface.” http://www.charmming.org.

[180] A. Tiwari and A. K. T. Sekhar, “Workflow based framework for life science informatics.,” *Comput Biol Chem*, vol. 31, pp. 305–319, Oct 2007.

[181] “Scitegic data analysis and reporting platform,” 2008. http://accelrys.com/products/scitegic/.

[182] “Inforsense: Agile, predictive & visual enterprise intelligence solutions,” 2008. http://www.inforsense.com.

[183] “Knime - konstanz information miner,” 2008. http://www.knime.org.

[184] T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, T. Carver, K. Glover, M. R. Pocock, A. Wipat, and P. Li, “Taverna: a tool for the composition and enactment of bioinformatics workflows.,” *Bioinformatics*, vol. 20, pp. 3045–3054, Nov 2004.

[185] “Condor. high thoughput computing,” 2008. http://www.cs.wisc.edu/condor/condorg/.

[186] “Discovering dengue drug-together,” 2008. http://www.utmb.edu/discoveringdenguedrugs%2Dtogether/.

[187] “Screensaver lifesaver - university of oxford.” http://www.chem.ox.ac.uk/curecancer.html.

[188] “Novartis pharmaceuticals,” 2006. http://www.novartis.com/index.shtml.

[189] “Distributed desktop grid, pc refresh help novartis enhance innovation,” 2003. http://www.univaud.com/about/resources/files/cs-novartis.pdf.

[190] K. Irwin, “Ucsf dock,” 2006. http://dock.compbio.ucsf.edu/.

[191] “European organisation for nuclear research.” http://public.web.cern.ch/public/.

[192] “The large hadron collider.” http://lhc.web.cern.ch/lhc/.

[193] “Light weight middleware for grid computing,” 2006. http://glite.web.cern.ch/glite/.

[194] N. Jacq, J. Salzemann, Y. Legre, M. Reichstadt, F. Jacq, M. Zimmermann, A. Maass, M. Sridhar, K. Vinod-Kusam, H. Schwichtenberg, M. Hofmann, and V. Breton, “Demonstration of in silico docking at a large scale on grid infrastructure.,” *Stud Health Technol Inform*, vol. 120, pp. 155–157, 2006.

[195] “Virtual organization membership service,” 2007. http://vdt.cs.wisc.edu/VOMS-documentation.html.

[196] “Auvergrid,” 2006. www.auvergrid.fr.

[197] “Eumedgrid,” 2006. www.eumedgrid.org.

[198] P. Pattanaik, J. Raman, and H. Balaram, “Perspectives in drug design against malaria.,” *Curr Top Med Chem*, vol. 2, pp. 483–505, May 2002.

[199] I. Y. Gluzman, S. E. Francis, A. Oksman, C. E. Smith, K. L. Duffin, and D. E. Goldberg, “Order and specificity of the plasmodium falciparum hemoglobin degradation pathway.,” *J Clin Invest*, vol. 93, pp. 1602–1608, Apr 1994.

[200] K. A. Kolakovich, I. Y. Gluzman, K. L. Duffin, and D. E. Goldberg, “Generation of hemoglobin peptides in the acidic digestive vacuole of plasmodium falciparum implicates peptide transport in amino acid production.,” *Mol Biochem Parasitol*, vol. 87, pp. 123–135, Aug 1997.

[201] J. B. Dame, C. A. Yowell, L. Omara-Opyene, J. M. Carlton, R. A. Cooper, and T. Li, “Plasmepsin 4, the food vacuole aspartic proteinase found in all plasmodium spp. infecting man.,” *Mol Biochem Parasitol*, vol. 130, pp. 1–12, Aug 2003.

[202] R. Banerjee, J. Liu, W. Beatty, L. Pelosof, M. Klemba, and D. E. Goldberg, “Four plasmepsins are active in the plasmodium falciparum food vacuole, including a protease with an active-site histidine.,” *Proc Natl Acad Sci U S A*, vol. 99, pp. 990–995, Jan 2002.

[203] S. E. Francis, D. J. Sullivan, and D. E. Goldberg, “Hemoglobin metabolism in the malaria parasite plasmodium falciparum.,” *Annu Rev Microbiol*, vol. 51, pp. 97–123, 1997.

[204] K. Ersmark, B. Samuelsson, and A. Hallberg, “Plasmepsins as potential targets for new antimalarial therapy.,” *Med Res Rev*, vol. 26, pp. 626–666, Sep 2006.

[205] G. H. Coombs, D. E. Goldberg, M. Klemba, C. Berry, J. Kay, and J. C. Mottram, “Aspartic proteases of plasmodium falciparum and other parasitic protozoa as drug targets.,” *Trends Parasitol*, vol. 17, pp. 532–537, Nov 2001.

[206] A. M. Silva, A. Y. Lee, S. V. Gulnik, P. Maier, J. Collins, T. N. Bhat, P. J. Collins, R. E. Cachau, K. E. Luker, I. Y. Gluzman, S. E. Francis, A. Oksman, D. E. Goldberg, and J. W. Erickson, “Structure and inhibition of plasmepsin ii, a hemoglobin-degrading enzyme from plasmodium falciparum.,” *Proc Natl Acad Sci U S A*, vol. 93, pp. 10034–10039, Sep 1996.

[207] K. Ersmark, I. Feierberg, S. Bjelic, E. Hamelink, F. Hackett, M. J. Blackman, J. Hultén, B. Samuelsson, J. Aqvist, and A. Hallberg, “Potent inhibitors of the plasmodium falciparum enzymes plasmepsin i and ii devoid of cathepsin d inhibitory activity.,” *J Med Chem*, vol. 47, pp. 110–122, Jan 2004.

[208] O. A. Asojo, E. Afonina, S. V. Gulnik, B. Yu, J. W. Erickson, R. Randad, D. Medjahed, and A. M. Silva, “Structures of ser205 mutant plasmepsin ii from plasmodium falciparum at 1.8 a in complex with the inhibitors rs367 and rs370.,” *Acta Crystallogr D Biol Crystallogr*, vol. 58, pp. 2001–2008, Dec 2002.

[209] “The zinc database,” 2005. http://blaster.docking.org/zinc/.

[210] J. J. Irwin and B. K. Shoichet, “Zinc–a free database of commercially available compounds for virtual screening.,” *J Chem Inf Model*, vol. 45, no. 1, pp. 177–182, 2005.

[211] “The leader in discovery chemitry services and screening libraries,” 2005. http://www.chembridge.com/.

[212] “Autodock,” 2005. http://autodock.scripps.edu/.

[213] M. Rarey, B. Kramer, and T. Lengauer, “The particle concept: placing discrete water molecules during protein-ligand docking predictions.,” *Proteins*, vol. 34, pp. 17–28, Jan 1999.

[214] B. Kramer, M. Rarey, and T. Lengauer, “Evaluation of the flexx incremental construction algorithm for protein-ligand docking.,” *Proteins*, vol. 37, pp. 228–241, Nov 1999.

[215] S. Jiang, S. T. Prigge, L. Wei, G. Ye, T. H. Hudson, L. Gerena, J. B. Dame, and D. E. Kyle, “New class of small nonpeptidyl compounds blocks plasmodium falciparum development in vitro by inhibiting plasmepsins.,” *Antimicrob Agents Chemother*, vol. 45, pp. 2577–2584, Sep 2001.

[216] G. Rastelli, G. Degliesposti, A. D. Rio, and M. Sgobba, “Binding estimation after refinement, a new automated procedure for the refinement and rescoring of docked ligands in virtual screening.,” *Chem Biol Drug Des*, vol. 73, pp. 283–286, Mar 2009.

[217] P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D. A. Case, and T. E. Cheatham, “Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models.,” *Acc Chem Res*, vol. 33, pp. 889–897, Dec 2000.

[218] K. P. Massova I, “Combined molecular mechanical and continuum solvent approach (mm-pbsa/gbsa) to predict ligand binding.,” *Perspect Drug Discov*, vol. 18, p. 113–135, 2000.

[219] J. Srinivasan, T. Cheatham, P. Cieplak, P. Kollman, and D. Case, “Continuum solvent studies of the stability of dna, rna, and phosphoramidate-dna helices,” *J. Am. Chem. Soc.,*, vol. 120, pp. 9401 – 9409, 1998.

[220] B. R. Brooks, D. Janezic, and M. K. Karplus, “Harmonic analysis of large systems i. methodology,” *J. Comput. Chem.*, vol. 16, pp. 1522–1542, 1995.

[221] D. Xie, S. Gulnik, L. Collins, E. Gustchina, L. Suvorov, and J. W. Erickson, “Dissection of the ph dependence of inhibitor binding energetics for an aspartic protease: direct measurement of the protonation states of the catalytic aspartic acid residues.,” *Biochemistry*, vol. 36, pp. 16166–16172, Dec 1997.

[222] E. Alexov, “Calculating proton uptake/release and binding free energy taking into account ionization and conformation changes induced by protein-inhibitor association: application to plasmepsin, cathepsin d and endothiapepsin-pepstatin complexes.,” *Proteins*, vol. 56, pp. 572–584, Aug 2004.

[223] P. Willett, J. M. Barnard, and G. M. Downs, “Chemical similarity searching,” *J. Chem. Inf. Comput. Sci*, vol. 38, pp. 983–996, 1998.

[224] E. S. Istvan and D. E. Goldberg, “Distal substrate interactions enhance plasmepsin activity.,” *J Biol Chem*, vol. 280, pp. 6890–6896, Feb 2005.

[225] K. E. Luker, S. E. Francis, I. Y. Gluzman, and D. E. Goldberg, “Kinetic analysis of plasmepsins i and ii aspartic proteases of the plasmodium falciparum digestive vacuole.,” *Mol Biochem Parasitol*, vol. 79, pp. 71–78, Jul 1996.

[226] J. Hill, L. Tyas, L. H. Phylip, J. Kay, B. M. Dunn, and C. Berry, “High level expression and characterisation of plasmepsin ii, an aspartic proteinase from plasmodium falciparum.,” *FEBS Lett*, vol. 352, pp. 155–158, Sep 1994.

[227] E. D. Matayoshi, G. T. Wang, G. A. Krafft, and J. Erickson, “Novel fluorogenic substrates for assaying retroviral proteases by resonance energy transfer.,” *Science*, vol. 247, pp. 954–958, Feb 1990.

[228] “Spread of malaria by p. vivax,” 2007. http://usachppm.apgea.army.mil/Documents/FACT/18-041-0107_Vivax_Malaria.pdf.

[229] K. Becker, L. Tilley, J. L. Vennerstrom, D. Roberts, S. Rogerson, and H. Ginsburg, “Oxidative stress in malaria parasite-infected erythrocytes: host-parasite interactions.,” *Int J Parasitol*, vol. 34, pp. 163–189, Feb 2004.

[230] G. Riganese, R. Cardoso, D. Daniels, C. Bruns, and J. Tainer, “Characterizations of the electrophile binding site and substrate binding mode of 26kda glutathione s-transferase form schistosoma japonicum,” *PROTEINS, Structure, Function and Genetics*, vol. 5, pp. 137–146, 2003.

[231] D. Sheenan, G. Meade, V. Foley, and C. Dowd, “Structure function and evolution of glutathione transferases, implication for classification of non-mammalian members of an ancient super family,” *Biochemistry Journal*, vol. 360, pp. 1–16, 2001.

[232] P. Srivastva, S. Puri, K. Kamboj, and V. Pandey, “Glutathione s-transferase activity in malaria parasites,” *Tropical Medicine and International Health*, vol. 4, pp. 251–254., 1999.

[233] V. L. Dubois, D. F. Platel, G. Pauly, and J. Tribouley-Duret, “Plasmodium berghei: implication of intracellular glutathione and its related enzyme in chloroquine resistance in vivo.,” *Exp Parasitol*, vol. 81, pp. 117–124, Aug 1995.

[234] L. H. Miller, D. I. Baruch, K. Marsh, and O. K. Doumbo, “The pathogenic basis of malaria.,” *Nature*, vol. 415, pp. 673–679, Feb 2002.

[235] K. Fritz-Wolf, A. Becker, S. Rahlfs, P. Harwaldt, R. H. Schirmer, W. Kabsch, and K. Becker, “X-ray structure of glutathione s-transferase from the malarial parasite plasmodium falciparum.,” *Proc Natl Acad Sci U S A*, vol. 100, pp. 13821–13826, Nov 2003.

[236] M. Perbandt, C. Burmeister, R. D. Walter, C. Betzel, and E. Liebau, “Native and inhibited structure of a mu class-related glutathione s-transferase from plasmodium falciparum.,” *J Biol Chem*, vol. 279, pp. 1336–1342, Jan 2004.

[237] R. T. Koehler, H. O. Villar, K. E. Bauer, and D. L. Higgins, “Ligand-based protein alignment and isozyme specificity of glutathione s-transferase inhibitors.,” *Proteins*, vol. 28, pp. 202–216, Jun 1997.

[238] E. Liebau, B. Bergmann, A. M. Campbell, P. Teesdale-Spittle, P. M. Brophy, K. Lüersen, and R. D. Walter, “The glutathione s-transferase from plasmodium falciparum.,” *Mol Biochem Parasitol*, vol. 124, no. 1-2, pp. 85–90, 2002.

[239] J. K. Baird, “Chloroquine resistance in plasmodium vivax.,” *Antimicrob Agents Chemother*, vol. 48, pp. 4075–4083, Nov 2004.

[240] Ahlm, Wiström, and Carlsson, “Chloroquine-resistant plasmodium vivax malaria in borneo.,” *J Travel Med*, vol. 3, p. 124, Jun 1996.

[241] M. D. Young and R. W. Burgess, “Pyrimethamine resistance in plasmodium vivax malaria.,” *Bull World Health Organ*, vol. 20, no. 1, pp. 27–36, 1959.

[242] M. Imwong, S. Pukrittakayamee, S. Looareesuwan, G. Pasvol, J. Poirreiz, N. J. White, and G. Snounou, “Association of genetic mutations in plasmodium vivax dhfr with resistance to sulfadoxine-pyrimethamine: geographical and clinical correlates.,” *Antimicrob Agents Chemother*, vol. 45, pp. 3122–3127, Nov 2001.

[243] P. Kongsaeree, P. Khongsuk, U. Leartsakulpanich, P. Chitnumsub, B. Tarnchompoo, M. D. Walkinshaw, and Y. Yuthavong, “Crystal structure of dihydrofolate reductase from plasmodium vivax: pyrimethamine displacement linked with mutation-induced resistance.,” *Proc Natl Acad Sci U S A*, vol. 102, pp. 13046–13051, Sep 2005.

[244] W. Sirawaraporn, T. Sathitkul, R. Sirawaraporn, Y. Yuthavong, and D. V. Santi, “Antifolate-resistant mutants of plasmodium falciparum dihydrofolate reductase.,” *Proc Natl Acad Sci U S A*, vol. 94, pp. 1124–1129, Feb 1997.

[245] Y. Yuthavong, “Basis for antifolate action and resistance in malaria.,” *Microbes Infect*, vol. 4, pp. 175–182, Feb 2002.

[246] G. Rastelli, W. Sirawaraporn, P. Sompornpisut, T. Vilaivan, S. Kamchonwongpaisan, R. Quarrell, G. Lowe, Y. Thebtaranonth, and Y. Yuthavong, “Interaction of pyrimethamine,

cycloguanil, wr99210 and their analogues with plasmodium falciparum dihydrofolate reductase: structural basis of antifolate resistance.," *Bioorg Med Chem*, vol. 8, pp. 1117–1128, May 2000.

[247] G. Rastelli, S. Pacchioni, and M. D. Parenti, "Structure of plasmodium vivax dihydrofolate reductase determined by homology modeling and molecular dynamics refinement.," *Bioorg Med Chem Lett*, vol. 13, pp. 3257–3260, Oct 2003.

[248] J. Yuvaniyama, P. Chitnumsub, S. Kamchonwongpaisan, J. Vanichtanankul, W. Sirawaraporn, P. Taylor, M. D. Walkinshaw, and Y. Yuthavong, "Insights into antifolate resistance from malarial dhfr-ts structures.," *Nat Struct Biol*, vol. 10, pp. 357–365, May 2003.

[249] "Modeller," 2006. http://salilab.org/modeller/.

[250] D. A. Case, T. E. C. III, T. Darden, H. Gohlke, R. Luo, K. M. M. Jr., A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods, "The amber biomolecular simulation programs," *Journal of Computational Chemistry*, vol. 26, no. 16, pp. 1668–1688, 2005.

[251] A. Jakalian, D. B. Jack, and C. I. Bayly, "Fast, efficient generation of high-quality atomic charges. am1-bcc model: Ii. parameterization and validation.," *J Comput Chem*, vol. 23, pp. 1623–1641, Dec 2002.

[252] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, "Development and testing of a general amber force field.," *J Comput Chem*, vol. 25, pp. 1157–1174, Jul 2004.

[253] "Ucsf chimera," 2006. http://www.cgl.ucsf.edu/chimera.

[254] "Molecular operating environment," 2006. http://www.chemcomp.com.

Appendix

Appendix I: Python script converting sybyl mol2 format to pdbq format

```
from os import listdir, system
import os.path
inp_pdb_dir = "/home/kasam/Desktop/plasmepsin_AutoDock/mol2"
out_pdb_dir = "/home/kasam/Desktop/plasmepsin_AutoDock/lig_pdbq/"
pdb_list = listdir(inp_pdb_dir)
for pdb in pdb_list:
pdbq = ".pdbq"
filename = os.path.splitext(pdb)[0]+pdbq
cmd = "/home/bio/groupshare/software/dist305/bin/i86Linux2/autotors A
+7.5 h m c r %s %s" %(os.path.join(inp_pdb_dir,pdb), os.path.join(out_pdb_dir,filename))
print " >>", cmd
system(cmd)
```

Appendix II: Python script converting protein pdb format to pdbq format

```
from os import listdir
# change these filenames
inp_pdb_dir = "/home/kasam/Desktop/AutoDock_today/pdb/"
out_pdb_dir = "/home/kasam/Desktop/AutoDock_today/pro_pdbq/"
#looking for all files in the directory
pdb_list = listdir(inp_pdb_dir)
# doing all the adt stuff
for pdb in pdb_list:
self.readMolecule('%s%s' %( inp_pdb_dir, pdb), log=0)
self.add_hGC("%s" %(pdb[:4]),
polarOnly=1, renumber=1, method='noBondOrder',
log=0)
self.addKollmanCharges("%s" %(pdb[:4]),
log=0)
self.writePDBQ("%s" %(pdb[:4]),
sort=0, log=0, filename='%s%sq' %(out_pdb_dir,
pdb), pdbRec=("ATOM', t,r)ansformed=0, bondOrigin=())
self.deleteMol("%s" %(pdb[:4]),
log=0)
```

Appendix III: Python script converting protein pdbq format to pdbqs format

```
from os import listdir, system
```

```
import os.path
inp_pdb_dir = "/home/kasam/Desktop/plasmepsin_AutoDock/pro_pdbq"
out_pdb_dir = "/home/kasam/Desktop/plasmepsin_AutoDock/pdbqs/"
pdb_list = listdir(inp_pdb_dir)
for pdb in pdb_list:
cmd = "/home/bio/groupshare/software/dist305/bin/i86Linux2/addsol %s %ss" %
(os.path.join(inp_pdb_dir,pdb), os.path.join(out_pdb_dir,pdb))
print " >>", cmd
system(cmd)
```

Appendix IV: Python script creating Grid parameter file (gpf)

```
from os import listdir, system
import os.path
lig_dir = "/home/kasam/Desktop/plasmepsin_AutoDock/lig_pdbq"
macro_dir = "/home/kasam/Desktop/plasmepsin_AutoDock/pdbqs"
lig_list = listdir(lig_dir)
for lig in lig_list:
lig = os.path.splitext(lig)[0]
cmd = "/home/bio/groupshare/software/dist305/share/mkgpf3 %s %s.pdbq %s %
s.pdbqs" %(lig_dir, lig, macro_dir, lig)
print " >>", cmd
system(cmd)
```

Appendix V: Python script creating docking parameter file (dpf)

```
from os import listdir, system
import os.path
lig_dir = "/home/kasam/Desktop/plasmepsin_AutoDock/lig_pdbq"
macro_dir = "/home/kasam/Desktop/plasmepsin_AutoDock/pdbqs"
lig_list = listdir(lig_dir)
for lig in lig_list:
lig = os.path.splitext(lig)[0]
cmd = "/home/bio/groupshare/software/dist305/share/mkdpf3_kasam %s %s.pdbq %s
%s.pdbqs" %(lig_dir, lig, macro_dir, lig)
print " >>", cmd
system(cmd)
```

Appendix VI: Genetic algorithm settings used in AutoDock DC

```
"ga_pop_size 50 # number of individuals in population\n" .
"ga_num_evals 250000 # maximum number of energy evaluations\n" .
"ga_num_generations 27000# maximum number of generations\n" .
```

```
"ga_elitism 1 # num. of top individuals that automatically survive\n" .
"ga_mutation_rate 0.02 # rate of gene mutation\n" .
"ga_crossover_rate 0.80 # rate of crossover\n" .
"ga_window_size 10 # num. of generations for picking worst individual\n" .
"ga_cauchy_alpha 0 # ~mean of Cauchy distribution for gene mutation\n" .
"ga_cauchy_beta 1 # ~variance of Cauchy distribution for gene mutation\n" .
"set_ga # set the above parameters for GA or LGA\n\n";LS
parameters
$outtext # Local Search (Solis and Wets) Parameters (for LS alone
and for LGA).
"sw_max_its 300 # number of iterations of Solis and Wets local search.
"sw_max_succ 4 # number of consecutive successes before changing rho .
"sw_max_fail 4 # number of consecutive failures before changing rho.
"sw_rho 1.0 # size of local search space to sample" .
"sw_lb_rho 0.01 # lower bound on rho.
"ls_search_freq 0.06 # probability of performing local search on an individual .
"set_psw1 # set the above pseudoSolis
and Wets parameters.
```

Appendix VII : Parameter settings for minimizing water molecules for AutoDock

```
minimization prot fix; 11.4.: OK
&cntrl
imin=1,
ncyc=300,maxcyc=300,
ntx=1, irest=0,
ntpr=10, cut=20,
ntb = 0,
ntr=1, restraint_wt=999.0, restraintmask="(!@H=)",
&end
eof
$AMBERHOME/exe/sander O –I sandermin o
$target\_out_min p
$target.top c
$target.crd r
$target\_min.crd ref
$target.crd
rm sandermin
#$AMBERHOME/exe/ambpdb p
$target.top < $target\_min.crd > $target\_min.pdb
```

Appendix VIII: Flexx batch script used for large-scale docking

```
# LOAD PARAMETERS
set LIGAND $(LIGAND_CMD)
set RECEPTOR $(RECEPTOR_CMD)
set PDB $(PDB_CMD)
set SITE $(SITE_CMD)
set PREDICT $(PREDICT_CMD)
set SURFACE $(SURF_CMD)

# SCENARIO PARAMETERS will overwrite defaults from config.dat
set verbosity  1
set place_particles 1

# read a single protein
SELINP $(protein) in proteins.txt $(protein_nr)
SELINP $(ligand) in ligands.txt $(ligand_nr)

# read protein
receptor                      # change into receptor menu
  read $(protein)               # read protein
  trihash all                  # init datastructure
end                          # leave receptor menu

# read 15 test ligands of kasam
FOR_EACH $(nr) FROMTO 1 14

  ligand
    read kasam $(nr)
  end

  docking                       # change into docking menu
    selbas a                    # automatic selection of base frag
    placebas 3                   # place base frag (triangle alg.)
    complex all                  # build up complex
    cluster % % %
    seloutp $(protein)_$(ligand)_$(lig_start)_sol a y   # redirect output to file
    listsol 10                                    # one line info / top rank
    seloutp $(protein)_$(ligand)_$(lig_start)_mat a y
    listmat 10                                     #lists scores of best 10 sol
    seloutp screen                                   # redirect output to file screen
  end

  ligand
    write $(protein)_$(ligand)_$(lig_start) y y 1-10 n  # merge them
```

```
 end

END_FOR

# read the real molecules
FOR_EACH $(nr) FROMTO $(lig_start) $(lig_end)

 ligand
   read $(ligand) $(nr)
 end

 docking                    # change into docking menu
   selbas a                 # automatic selection of base frag
   placebas 3                 # place base frag (triangle alg.)
   complex all                # build up complex
   cluster % % %
   seloutp $(protein)_$(ligand)_$(lig_start)_sol a y   # redirect output to file
   listsol 10                                  # one line info / top rank
   seloutp $(protein)_$(ligand)_$(lig_start)_mat a y
   listmat 10                                   #lists scores of best 10 sol
   seloutp screen                                 # redirect output to file screen
 end
 ligand
   write $(protein)_$(ligand)_$(lig_start) y y 1-10 n  # merge them
 end
END_FOR
```

Appendix IX: List of all the software used in the current project.

Docking software: FlexX 2.0, FlexX-Pharm, FlexV (www.biosolveit.de) AutoDock 3.05, ADT (www.scripps.edu), Amber9

Others: Chimera, Rasmol, Babel, PDBkabsch, PDBtransform, Corina, Jcsearch, mdraw, ISIS draw, MOE, Tripos, VS explorer

Computational Grid infrastructure: EGEE

Scripts: Bash shell scripting, Python7 and Perl.

Appendix IX: WISDOM production environment for deploying docking and molecular dynamics applications

WISDOM production environment can be downloaded from:

https://sourceforge.net/projects/WISDOM-pe

Appendix X: Analysis of wild type PfDHFR results after molecular dynamics simulations

ANALYSIS OF DHFR MD results

Very good	Good	Bad	Known compound								
Rank	ZINC ID	Pbtot score	Active site after docking?	Active site after MD?	H Bonds with key residues				Orientation compared to WR9	Mobility afer MD	Comments
					ASP 54	NDP	H20	Others			
1	ZINC02924841	WRONG ENERGY									
2	ZINC03930026	WRONG ENERGY									
3	ZINC03932090	-59.79	Yes	Yes	1	0	1	0	Good	Low for the portion inside the protein	Portion outside the protein
4	ZINC04088977	-56.04	Yes	Yes					Not good		Very big molecule
5	ZINC02008198	-51.56	Yes	Yes	1	0	0	1HB SER 111 1HB GLY 64	Good	Low	Small molecule Check protonation
6	ZINC01554249	-50.04	Yes	Yes	2	0	0	1HB THR 185 1HB ILE 164	Not good	Low	S02 group Hydroxyl group
7	ZINC03823353	-48.7	Yes	Yes	1	0	1	0	Good	Low	
8	ZINC06581531	-48.21	Yes	Yes	2	0	1	1HB ILE 164 1HB ILE 40	Good	Low	Sulfide bridge Ether
9	ZINC05007934	-48.15	Yes	Yes	2	1	1	0	Good	Low	NO2 and SO2 groups which can be removed
10	ZINC02008201	-48.12	Yes	Yes	1	1	0	2HB ARG 122 1HB LYS 34	Good	Low	Check protonation
11	ZINC06040655	-48	Yes	Yes	0	0	0	0	Moderate	Low	Aromatic ring which can not make interaction with asp 54
12	ZINC04393111	-46.86	Yes	Yes	1	0	0	1HB LYS 34	Good	Low	
13	ZINC03919086	-45.98	Yes	Yes	1	0	0	1HB SER 111 1HB ARG 122	Good	Low	Stacking interaction with PHE 116
14	ZINC04896176	-45.93	Yes	No	1	0	0	1HB ILE 164 1HB SER 111	Good	High	Interaction SER 111: the groups are orthogonal
15	ZINC03989819	-45.9	Yes	Yes	2	0	1	1HB ILE 164 1HB ARG 122	Good	Low	
16	ZINC05579381	-45.77	Yes	Yes	0	0	0	1HB ARG122	Good	Low	Interaction SO2 with ARG122
17	ZINC03798623	-45.56	Yes	Yes	1	0	0	1HB ILE 164 2HB ARG 122	Not good	Low	
18	ZINC03937555	-45.49	Yes	Yes	2	0	1	1HB VAL 45	Good	Low	
19	ZINC05579385	-45.34	Yes	Yes	0	0	0	0			Very big molecule

Figure 42: Result analysis of wild type PfDHFR after molecular dynamics simulations. Virtual screening by molecular docking was performed at LPC-IN2P3 (by me). The molecular dynamics simulations were completely performed by Prof. Giulio Rasteli group at university of Modena. The results described in the table are combined effort of Prof. Giulio Rasteli team and LPC-IN2P3 (me).

Curriculum Vitae

Vinod Kumar Kasam
Colmantstr. 26,
53115, Bonn,
Germany

Date-of-Birth:	10-05-1976
Marital status:	Married
Nationality:	India

Work Experience

Since 12/2007	**Wissenschaftlicher Mitarbeiter,** Fraunhofer-Institute for algorithms and Scientific Computing (SCAI), Department of Bioinformatics, Schloss Birlinghoven, D-53754, Sankt Augustin , Germany
06/2006 – 11/2007	**Research Engineer,** Laboratoire de Physique Corpusculaire, CNRS-IN2P3, 24, avenue des Landais, 631177, AUBIERE Cedex , France
11/2002 - 11/2003	**Research Associate,** GVK Biosciences, Kundanbagh, Hyderabad, India

Academic Qualifications

11/2003 – 11/2005	Master of Science in Biology with biomedical sciences, University of Applied Sciences, Bonn, Germany
06/1997 – 10/2000	Master of Science in Biochemistry, Kakatiya University, Warangal, India,
06/1994 – 06/1997	Bachelor of Science in Chemistry, Kakatiya University, Warangal, India, 1997