DIPLOMARBEIT

Titel der Diplomarbeit

„Comparison of human, mouse and rat ABC transporters on basis of their substrate and inhibitor profile"

Verfasserin

Victoria Slubowski

angestrebter akademischer Grad

Magistra der Pharmazie (Mag.pharm.)

Wien, 2012

Studienkennzahl lt. Studienblatt:	A 449
Studienrichtung lt. Studienblatt:	Diplomstudium Pharmazie
Betreuer:	Univ.-Prof. Dr. Gerhard Ecker

Diese Arbeit widme ich meiner Mama, Krystyna Slubowski.

Danksagung

Mein besonderer Dank gilt Univ.-Prof. Dr. Gerhard F. Ecker für die Möglichkeit die Diplomarbeit in der Pharmacoinformatic Research Group durchführen zu dürfen. Trotz seines dichten Terminkalenders und vielerlei Verpflichtungen hatte er immer Zeit für die Beantwortung meiner Fragen. Überdies verstand er es die richtigen Rahmenbedingungen für ein angenehmes Arbeitsklima in der Gruppe zu schaffen.

Für die Unterstützung durch die Arbeitsgruppe möchte ich insbesondere Daniela, Freya, Kathi, Andi, Kamil, Nathan und Rene danken.

Des Weiteren möchte ich mich bei meinen Eltern, Krystyna und Andrzej Slubowski für ihre immerwährende Unterstützung in jeglicher Situation bedanken.

Nicht zuletzt würde ich mich gerne bei meinem Freund Lars bedanken, der mir von Anfang an immer mit Rat und Tat zu Seite gestanden hat.

Table of Contents

Abstract…………………………………………………………………………………...1

Zusammenfassung…………………………………………………………………….....3

I. Introduction 5
  A. Drug Development 5
  B. Biological and Pharmacological Background 12
    1. P-glycoprotein 12
  C. Computational Background 20
    1. Molecular Modeling 20
    2. Sequence Alignment 20
    3. Homology Modeling 24
    4. Docking 29
II. Aim of the study 35
III. Materials and Methods 37
  A. Multiple sequence alignment 37
  B. Sequence Alignment and Variability 43
  C. Homology Model 43
  D. Database Search 43
  E. Docking 43
IV. Results and Discussion 45
  A. Sequence Alignment 45
  B. Sequence Variability 46
  C. Homology Model 47
  D. Database search 51
    1. Transporter Database 51
    2. ChEMBL Database 53
    3. PubMed 55
  E. Docking 57
V. Summary and Outlook 61

Abstract in English

In the drug development process the ADME profile and the toxicity of a drug candidate are of major importance for its success. Here the ATP-dependant efflux pump P-glycoprotein (P-gp) plays a key role as it is expressed in biological barriers, like intestinal epithelium, blood brain barrier, proximal renal tubular cells and hepatocytes. The protein transports xenobiotic compounds with a broad substrate and inhibitor specificity out of the cell. Consequently, predictive *in silico* models for human P-gp activity are valuable tools in drug development.
However, at an early stage of drug development essential data is acquired in animal studies and consequently it is of utmost importance that drug candidates show a preferable pharmacokinetic and toxicity profile in animals. Thus next to existing predictive *in silico* models against human P-gp activity, predictive *in silico* models against rat and mouse would enable the avoidance of an early attrition in the following preclinical phase of animal in vivo studies. .

Recently a crystal structure of mouse P-gp was established and provides new possibilities for structure-based drug design approaches. The high sequence identity between rat and mouse P-gp (92%) and the importance of rats in animal ADME models motivated us to create a homology model of rat P-gp taking the crystallized mouse P-gp as a template. A multiple sequence alignment was performed using ClustalW2 and the resulting alignment was then used within MODELLER for model generation.

Subsequently the docking software GOLD was used to dock 6 rat P-gp inhibitors with known IC_{50} values into the rat homology model. Docking poses were analyzed and showed frequent interactions between the ligand poses and F70 (TM helix 1) and F335 (TM helix 6). Also residue T306 (TM helix 5) was involved, whose human analogue T307 was (experimentally) shown to be important in ligand interactions. The predictive power of the model could be validated by comparing the rankings resulting from the scoring function GOLDScore and the experimentally determined activity: the docking was able to correctly assign the ranking for all but one of the experimentally tested compounds (only ranks 3 and 4 were switched).

Zusammenfassung

Während des Drug Development Prozesses sind ADME und Toxizität eines Wirkstoff Kandidaten ausschlaggebend für seinen Erfolg. Hier spielt die ATP-abhängige Efflux Pumpe P-Glykoprotein (P-gp), die im Darmepithel, in der Blut-Hirn-Schranke, in proximalen Tubuluszellen der Niere und in Hepatozyten exprimiert wird, eine entscheidende Rolle. Das Protein transportiert xenobiotische Substanzen mit einem breiten Substrat und Inhibitor Profil aus der Zelle. Folglich sind vorhersagende *in silico* Modelle für das humane P-gp ein wertvolles Instrument im Drug Development. Allerdings werden in einem frühen Stadium des Drug Developments essentielle Daten in Tierstudien gewonnen, deswegen ist es besonders wichtig, dass ein Wirkstoffkandidat ein günstiges pharmakokinetisches Profil im Tiermodell zeigt. Daher könnten, abgesehen von den bestehenden *in silico* Vorhersagemodellen für das menschliche P-gp, *in silico* Vorhersagemodelle für das Ratten und Maus P-gp die Abbruchrate in den in vivo Tierstudien während der nächsten präklinischen Phase senken.

Kürzlich wurde die Kristallstruktur des Maus P-gp aufgeklärt und schafft somit neue Möglichkeiten für strukturbasiertes Wirkstoffdesign. Die hohe Sequenzidentität zwischen Ratte und Maus (92%) und die Bedeutung von Ratten in ADME Tiermodellen motivierte uns ein Homologie Modell der Ratte zu entwickeln, als dessen Vorlage die Kristallstruktur des Maus P-gp herangezogen wurde. Ein multiples Sequenzalignment wurde mit ClustalW2 durchgeführt und das resultierende Alignement wurde für die Modellberechnung mit MODELLER eingesetzt.

Anschließend wurden 6 P-gp Inhibitoren der Ratte mit bekannten IC_{50} Werten mit Hilfe der Docking Software GOLD in das Ratten Homologie Modell gedockt. Die Analyse der Docking Posen zeigte häufige Interaktionen zwischen den Aminosäuren F70 (TM Helix 1) und F335 (TM Helix 6). Auch Aminosäure T306 war an Interaktionen beteiligt, dessen humanes Analogon T307 (experimentell) nachweislich bei Ligandeninteraktionen von Bedeutung ist. Die Vorhersagekraft des Modells konnte durch Vergleich der Ranking Ergebnisse, die mit Hilfe von GOLDScore berechnet wurden, mit den experimentell getesteten Aktivitäten validiert werden: das Docken war in der Lage alle außer einer experimentell getesteten Verbindungen richtig zuzuordnen (nur Nummer 3 und 4 waren vertauscht).

I. Introduction

A. Drug Development

The drug development process takes about 10-15 years to develop a new drug from the discovery until the chance of treating patients. The costs of research and development of each successful drug are in average approximately 800 million to 1 billion U.S. dollars. The failures are included in that amount: for every 5000-10000 compounds which enter the research and development pipeline in the end only one is approved [**Figure 1**] [1-4].

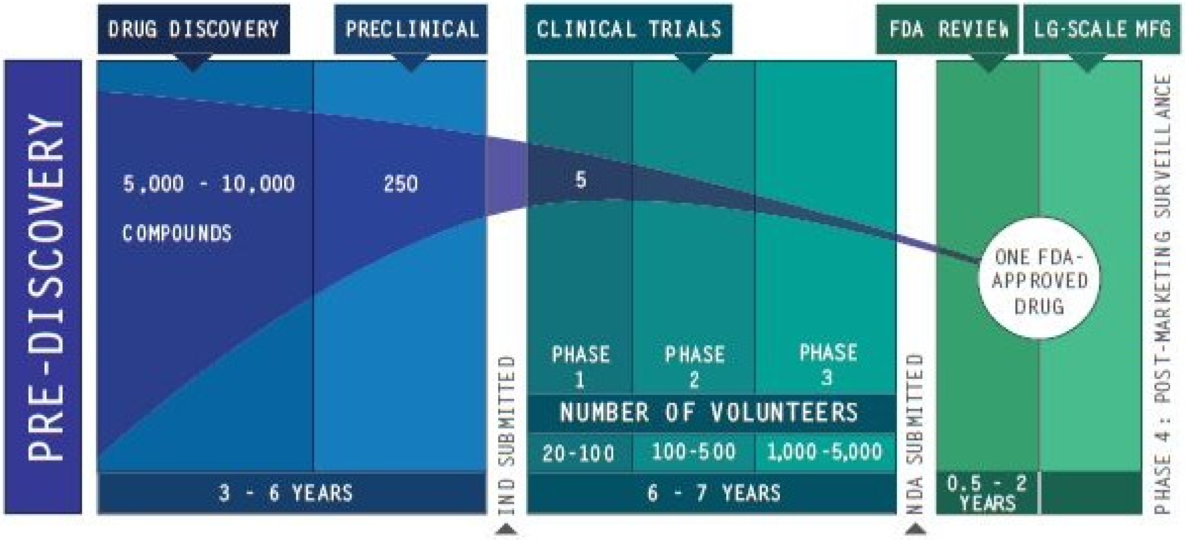

DRUG DISCOVERY
PRECLINICAL
CLINICAL TRIALS
FDA REVIEW
LG-SCALE MFG
PRE-DISCOVERY
5,000 - 10,000
COMPOUNDS
250
5
ONE FDA-APPROVED DRUG
PHASE 1
PHASE 2
PHASE 3
NUMBER OF VOLUNTEERS
20-100
100-500
1,000-5,000
3 - 6 YEARS
IND SUBMITTED
6 - 7 YEARS
NDA SUBMITTED
0.5 - 2 YEARS
PHASE 4: POST-MARKETING SURVEILLANCE

***Figure 1** – Depiction of the drug development process, figure taken from [3].*

Identify Disease

Before starting the discovery of a new potential drug, the focus lies on understanding the underlying disease and the cause of the condition for finding the best possible treatment. It is of great importance to understand how the genes are altered, how that affects the proteins they encode and how those proteins interact with each other in living cells, how those affected cells change the specific tissue they are in, and finally how the disease affects the entire patient. This knowledge is the fundament for an accurate treatment of the disease.

However, the research and the development of new drugs often end in nowhere. Even if your research is successful there are still many years of work and possible dead ends ahead before the understanding of a disease can be turned into a new treatment.

Target Identification

The next step after understanding the disease and its cause is to choose the target for a potential new drug. A target most often is a single macromolecule, e.g. a gene or protein, which plays a role in the explored disease. At this point of research it plays an important role to select such a target that is able to interact and to be modified by a new drug molecule (drugable target).

Target Validation

Once a potential target is chosen it has to be shown that it is involved in the disease and can be accessed and also affected by a drug. This step of research is essential to avoid promising looking drug candidates to finish in dead ends.

Drug Discovery

After having understood the disease and having found a validated target the search for a drug starts. The focus lies on a molecule, which can also be called “lead compound” that is able to interact with the target in a way to modify the course of the disease. If this search turns out to be successful in many years and after a lot of testing the lead compound could become a new drug.

There are different ways to determine the lead compound:

1. From nature: In former times there were no high tech methods to find new compounds in the way we discover them today. In the lack of these techniques nature very often delivered templates for new drugs, e.g. antibiotics. There are surely still a lot more drugs we can copy from nature.
2. De novo: The big progress in natural sciences makes it today even possible to design molecules from scratch. Computer modeling can be used to find out what kinds of molecules could have an effect on the target.
3. High-throughput Screening: This method is mostly used to discover hits which then might evolve to lead compounds. A hit is generated by a yes or no question; a lead is a

hit which is selected for further studies. The progress in robotics and computational power allows testing billions of compounds against the target to check whether any compound could be active. After the evaluation of the results some of the tested compounds are chosen for further studies.

4. Biotechnology: Another possibility to find new lead compounds is to genetically design living cells which produce disease-fighting biological molecules.

Early Safety Profiling

Lead compounds are tested with numerous tests to evaluate in an early stage of drug development the safety of the possible new drug. **A**bsorption, **D**istribution, **M**etabolism, **E**xcretion (ADME) and toxicological parameters, in short pharmacokinetics, are evaluated for each compound.

Drugs have to be:

- absorbed into the blood,
- distributed to the proper organ,
- metabolized efficiently and effectively,
- successfully excreted from the body and
- shown to be not toxic.

Pharmacokinetic tests can be performed in living cells or in animals (preferably mouse or rat)

Lead Optimization

After the screening and the first safety tests lead compounds are modified in different ways to find more effective and safer derivatives. Various properties can be changed in the molecule to make it more hydrophilic, lipophilic, acidic, basic, etc. The newly generated derivatives are tested and out of the test results further changes can be done to step by step develop molecules with even better properties. In the end a potential drug candidate is received.

Already at this point of the drug development the formulation, the delivery mechanism and the large-scale production of the new drug should be thought of.

What kind of inactive substances could be possibly used?

How should the new drug be assembled to dissolve at the right place and time?

Is it going to be an oral drug, an injection, an inhalation, etc.?

Is it possible to produce the new drug in large quantities?

All these questions should already be answered at this point of the development.

Pre-clinical Tests

In pre-clinical tests the one or more optimized lead compounds are tested more extensively to establish whether the drugs are safe to be tested in humans. For this purpose in vitro and in vivo tests are performed. In vitro is Latin and means “in glass”. As the name already indicates the experiments are performed in test tubes, Petri dishes or beakers. The expression in vivo is as well Latin and has the meaning “in life”. The in vivo tests are carried out in living animals. The purpose of these experiments is to interpret the mode of action of the drug and its safety. To acquire the approval for studies in humans the requirements to a drug candidate are extremely high.

In this phase it is necessary to give again the technological aspects some thoughts as well. The production of a larger quantity of the drug for a possible upcoming clinical trial needs to be planned precisely. The translation from a smaller to a larger production is not that easily performed. If the drug would be approved even another scale up would become necessary.

At this point already several years have passed and a lot of different studies have been performed. From the originally 5000 to 10000 compounds only one to five molecules are left in the development process. Next they are going to be tested in clinical trials.

IND Application

In US, an Investigational New Drug (IND) application has to be submitted to the United States Food and Drug Administration (FDA) before any clinical trial can be started. The contents of the application is supposed to contain the results of the preclinical studies, the candidate drug’s chemical structure, its mode of action in the body, a list of side effects and manufacturing information. Further a detailed plan of the clinical trial explaining how, where and by whom the studies will be carried out must be included.

The major concern of the FDA is the health of the participants of the clinical trial. All possible risks have to be ruled out in advance.

The trial is observed continuously and can be stopped by the FDA or the sponsor company at any time if problems occur. In contrast, it is as well possible to stop a trial and put the compound immediately to the market, because the drug is acting so well that it would be unethical to hold it back from other patients.

During the ongoing clinical trial the sponsor company is obliged to report regularly to the FDA.

Clinical Phase I

In Phase I the drug candidate is tested in healthy volunteers for the first time. Usually 20 to 100 patients are chosen for this purpose. The focus of the Phase I trial lies on finding out whether the drug candidate is safe in humans or not.

The following questions concerning pharmacokinetics and pharmacodynamics are of interest:

- How is the drug absorbed, metabolized and eliminated from the body?
- Are there any side effects?
- Do we experience all main desirable effects?

With the help of the answers to these questions it can be determined if the drug candidate should be further developed and if yes, what dosing range is safe.

Clinical Phase II

In the Phase II trial the drug's effectiveness is tested in around 100 to 500 patients. The volunteers suffer from the studied disease. In this phase of the study the short term side effects and risks of the drug candidate are tested.

Further interest in the Phase II lies in the following tasks:

Is the working mechanism the expected one?

Does the condition improve?

What dosage and schedule for drug use is optimal?

If after all the results still look promising, the much larger Phase III trial needs to be prepared.

Clinical Phase III

In Phase III the focus lies on generating statistically significant data about safety, efficacy and the overall benefit-risk-relationship of the drug. Therefore a much higher number of patients is needed (around 1000 to 5000). The most important aspect of this phase is the determination if the drug is safe and effective. Additionally the basis for labeling instructions, like information on interactions with other medicines, is provided to ensure the right use of the drug.

The Phase III trial is the most expensive and longest phase of all. Numerous different sites around the world usually participate in Phase III to ensure a large transverse profile of different patients. The management of all sites and the interpretation of their results and data is a huge challenge.

Throughout the Phase III trial other serious issues should be resolved as well. The full scale production of the new drug is a critical step and requires to be planned in every detail.

However before this can become reality a sophisticated application for FDA approval ought to be prepared.

FDA Approval

As soon as all 3 phases are finished the data is evaluated by the sponsoring company. If the data again confirm that the new drug is safe and effective the company submits a New Drug Application (NDA) which may consist of 100.000 pages. The FDA has to decide if the drug can be approved to the market. The NDA contains all results from the previous years and suggestions for manufacturing and labeling of the new drug.

The application is reviewed by FDA experts who have to decide if the drug is safe and effective enough to be approved. Therefore the risk-benefit-ratio is consulted, the package insert is checked for every needed information and the methods to produce the drug have to guarantee its quality. When all these aspects are positive the FDA approves the drug. In contrast, the FDA might request more information before an approval can be given or deny the approval right away.

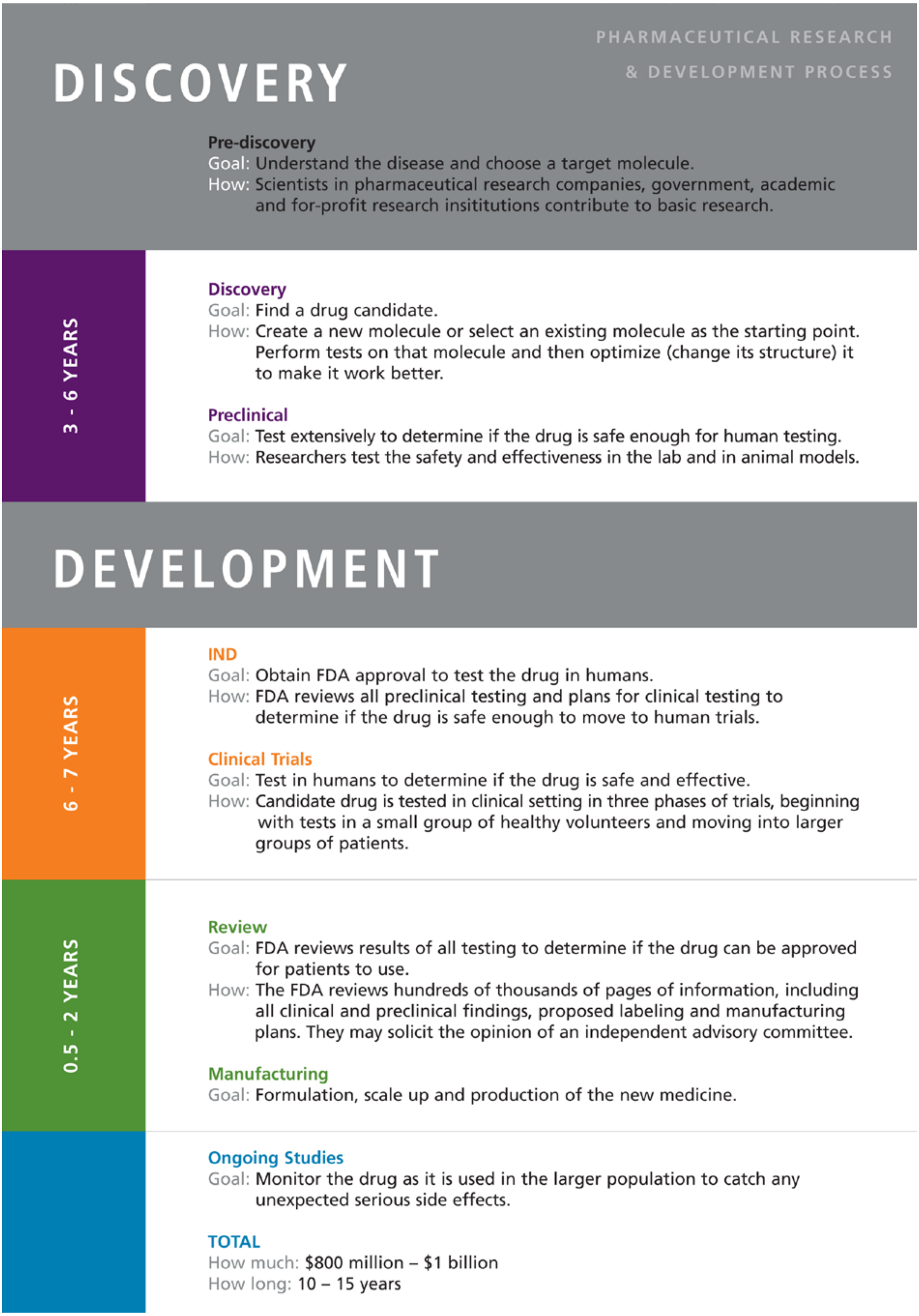
Manufacturing

The step from small scale to large scale manufacturing is a major undertaking. In many cases new manufacturing facilities must be built or old ones reconstructed because the manufacturing process varies from drug to drug. The FDA requires from each facility to follow the guidelines for Good Manufacturing Practices (GMP).

Ongoing Studies and Phase 4 Trials

Even after the approval the research on a new drug doesn’t stop. With the larger number of patients taking the drug the company is obliged to submit reports regularly, as well as cases of adverse drug reactions to the FDA.

Additionally sometimes further studies are required by the FDA even on an already approved drug. They are called Phase IV trials. The purpose of these studies can be the evaluation of long term safety or the affects of the drug on a specific subgroup of patients [1-4] [**Figure 2**].



PHARMACEUTICAL RESEARCH & DEVELOPMENT PROCESS

DISCOVERY

Pre-discovery
Goal: Understand the disease and choose a target molecule.
How: Scientists in pharmaceutical research companies, government, academic and for-profit research insititutions contribute to basic research.

3 - 6 YEARS

Discovery
Goal: Find a drug candidate.
How: Create a new molecule or select an existing molecule as the starting point. Perform tests on that molecule and then optimize (change its structure) it to make it work better.

Preclinical
Goal: Test extensively to determine if the drug is safe enough for human testing.
How: Researchers test the safety and effectiveness in the lab and in animal models.

DEVELOPMENT

6 - 7 YEARS

IND
Goal: Obtain FDA approval to test the drug in humans.
How: FDA reviews all preclinical testing and plans for clinical testing to determine if the drug is safe enough to move to human trials.

Clinical Trials
Goal: Test in humans to determine if the drug is safe and effective.
How: Candidate drug is tested in clinical setting in three phases of trials, beginning with tests in a small group of healthy volunteers and moving into larger groups of patients.

0.5 - 2 YEARS

Review
Goal: FDA reviews results of all testing to determine if the drug can be approved for patients to use.
How: The FDA reviews hundreds of thousands of pages of information, including all clinical and preclinical findings, proposed labeling and manufacturing plans. They may solicit the opinion of an independent advisory committee.

Manufacturing
Goal: Formulation, scale up and production of the new medicine.

Ongoing Studies
Goal: Monitor the drug as it is used in the larger population to catch any unexpected serious side effects.

TOTAL
How much: $800 million – $1 billion
How long: 10 – 15 years

Figure 2 *– Scheme of the pharmaceutical research and development process, Figure taken from [3].*

B. Biological and Pharmacological Background

1. P-glycoprotein

P-glycoprotein (P-gp) [**Figure 3**] is a protein in the cell membrane of eukaryotes and prokaryotes. P-gp is able to transport a wide variety of substrates against a concentration gradient out of the cell using adenosine triphosphate (ATP) as energy supplier which is bound and hydrolyzed at the P-gps nucleotide binding domain (NBD). The ATP-dependant efflux pump is extensively expressed in the intestinal epithelium, hepatocytes, renal proximal tubular cells, adrenal gland, capillary endothelial cells and blood brain barrier.

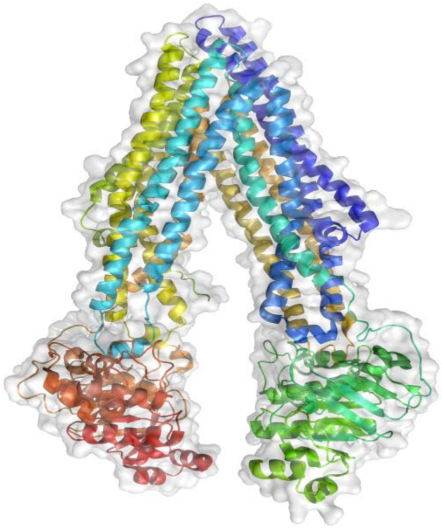


***Figure 3** – Crystallographic structure of the mouse mdr3 protein.*

P-gp belongs to the ABC-transporter family and further to the Multidrug-Resistance-Protein subfamily. Therefore it is also called ATP-binding cassette sub-family B member 1 (ABCB1) or Multidrug-Resistance-Protein 1 (MDR1).

ABC-transporters form a large group of transmembrane proteins that all have an ATP binding cassette (ABC) domain in common. They can be divided into three main functional groups: importers, exporters and proteins involved in gene expression regulation and DNA repair [5]. In prokaryotes importers are responsible for the mediation of nutrients into the cell. They do not exist in eukaryotes. Exporters are present in prokaryotes as well as in eukaryotes and mediate the efflux of xenobiotic compounds.

The human genome carries 49 ABC genes, set in seven subfamilies and named A to G [6] as shown in **Table 1 and Figure 4**. ABC-transporters are involved in a large variety of physiological processes, consequently they play a role in numerous diseases, e.g. tumor resistance, cystic fibrosis, bacterial multidrug resistance and other inherited human diseases as well.

Table 1 *- Human ABC transporter genes, chromosomal location, number of exons and their functions [6]*

Gene	*Chromosome location*	*Exons*	*Function*
ABCA1	*9q31.1*	*36*	*Cholesterol efflux onto HDL*
ABCA2	*9q34*	*27*	*Drug resistance*
ABCA3	*16p13.3*	*26*	*Multidrug resistance*
ABCA4	*1p22*	*38*	*N-retinylidene-phosphatidylethanolamine(PE) efflux*
ABCA5	*17q24.3*	*31*	*Urinary diagnostic marker for prostatic intraepithelial neoplasia (PIN)*
ABCA6	*17q24.3*	*35*	*Multidrug resistance*
ABCA7	*19p13.3*	*31*	*Cholesterol efflux*
ABCA8	*17q24*	*31*	*Transports certain lipophilic drugs*
ABCA9	*17q24.2*	*31*	*Might play a role in monocyte differentiation and macrophage lipid homeostasis*
ABCA10	*17q24*	*27*	*Cholesterol-responsive gene*
ABCA12	*2q34*	*37*	*Has implications for prenatal diagnosis*
ABCA13	*7p12.3*	*36*	*Inherited disorder affecting the pancreas*
ABCB1	*7q21.1*	*20*	*Multidrug resistance*
ABCB2	*6p21.3*	*11*	*Peptide transport*
ABCB3	*6p21.3*	*11*	*Peptide transport*

Gene	*Chromosome location*	*Exons*	*Function*
ABCB4	*7q21.1*	*25*	*Phosphatidylcholine (PC) transport*
ABCB5	*7p15.3*	*17*	*Melanogenesis*
ABCB6	*2q36*	*19*	*Iron transport*
ABCB7	*Xq12-q13*	*14*	*Fe/S cluster transport*
ABCB8	*7q36*	*15*	*Intracellular peptide trafficking across membranes*
ABCB9	*12q24*	*12*	*Located in lysosomes*
ABCB10	*1q42.13*	*13*	*Export of peptides derived from proteolysis of inner-membrane proteins*
ABCB11	*2q24*	*26*	*Bile salt transport*
ABCC1	*16p13.1*	*31*	*Drug resistance*
ABCC2	*10q24*	*26*	*Organic anion efflux*
ABCC3	*17q22*	*19*	*Drug resistance*
ABCC4	*13q32*	*19*	*Nucleoside transport*
ABCC5	*3q27*	*25*	*Nucleoside transport*
ABCC6	*16p13.1*	*28*	*Expressed primarily in liver and kidney*
ABCC7	*7q31.2*	*23*	*Chloride ion channel (same as CFTR gene in cystic fibrosis)*
ABCC8	*11p15.1*	*30*	*Sulfonylurea receptor*
ABCC9	*12p12.1*	*32*	*Encodes the regulatory SUR2A subunit of the cardiac K^+(ATP) channel*
ABCC10	*6p21.1*	*19*	*Multidrug resistance*
ABCC11	*16q12.1*	*25*	*Drug resistance in breast cancer*
ABCC12	*16q12.1*	*25*	*Multidrug resistance*
ABCC13	*21q11.2*	*6*	*Encodes a polypeptide of unknown function*
ABCD1	*Xq28*	*9*	*Very-long-chain fatty acid (VLCFA) transport*
ABCD2	*12q11–q12*	*10*	*Major modifier locus for clinical diversity in X-linked ALD (X-ALD)*
ABCD3	*1p22–p21*	*16*	*Involved in import of fatty acids and/or fatty acyl-coenzyme As into the peroxisome*
ABCD4	*14q24*	*19*	*May modify the ALD phenotype*
ABCE1	*4q31*	*14*	*Oligoadenylate-binding protein*
ABCF1	*6p21.33*	*19*	*Susceptibility to autoimmune pancreatitis*

Gene	*Chromosome location*	*Exons*	*Function*
ABCF2	*7q36*	*14*	*Tumour suppression at metastatic sites and in endocrine pathway for breast cancer/drug resistance*
ABCF3	*3q27.1*	*21*	*Also present in promastigotes (one of five forms in the life cycle of trypanosomes)*
ABCG1	*21q22.3*	*13*	*Cholesterol transport*
ABCG2	*4q22*	*16*	*Toxicant efflux, drug resistance*
ABCG4	*11q23.3*	*15*	*Found in macrophage, eye, brain and spleen*
ABCG5	*2p21*	*11*	*Sterol transport*
ABCG8	*2p21*	*10*	*Sterol transport*

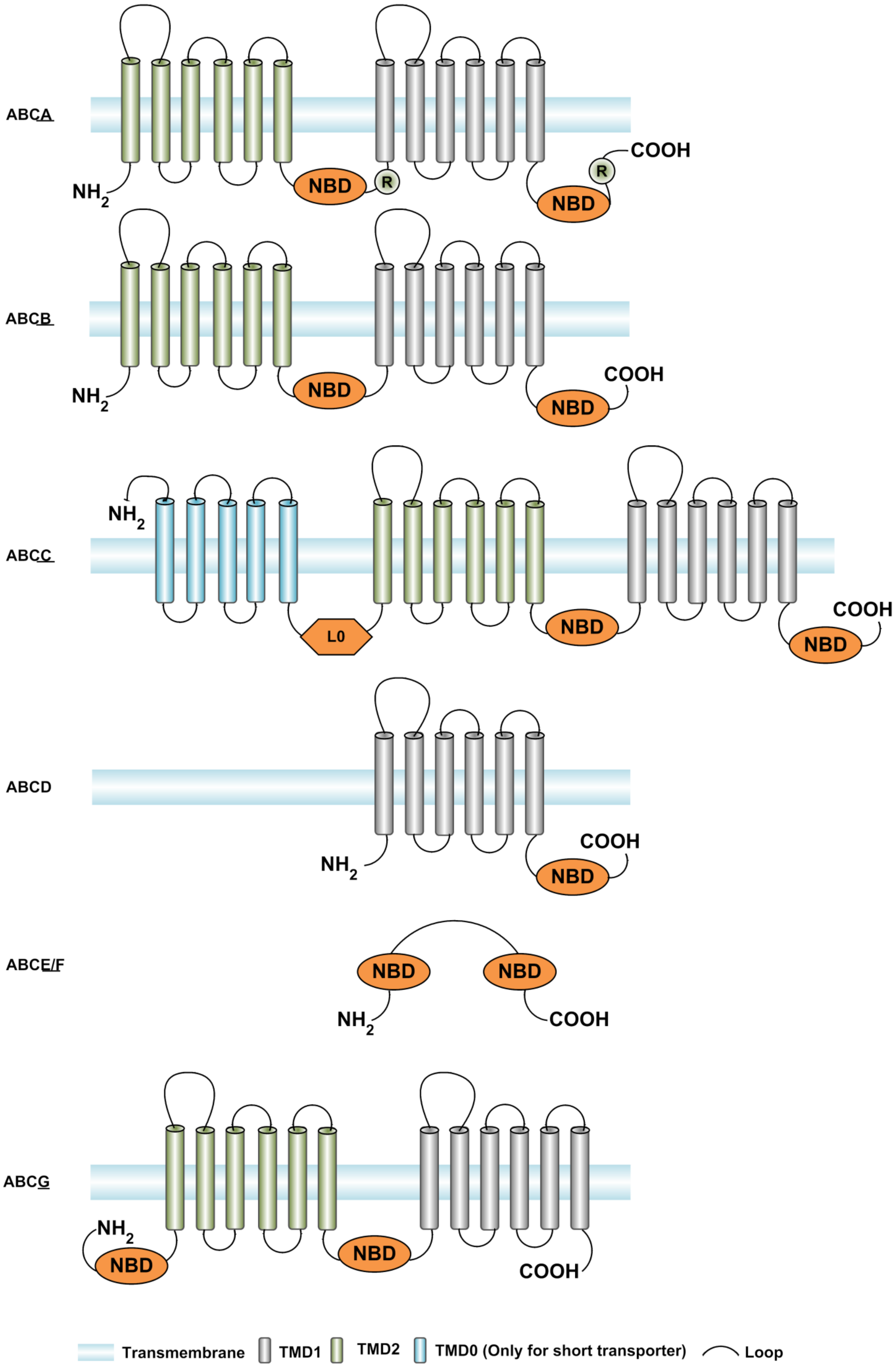

ABCA
NH2
NBD
R
NBD
R
COOH
ABCB
NH2
NBD
NBD
COOH
ABCC
NH2
L0
NBD
NBD
COOH
ABCD
NH2
NBD
COOH
ABCE/F
NBD
NBD
NH2
COOH
ABCG
NH2
NBD
NBD
COOH
Transmembrane
TMD1
TMD2
TMD0 (Only for short transporter)
Loop

Figure 4 *– Schematic depiction of the transmembrane domains of ABC subfamilies A to G.*

In 2009 the mouse P-gp crystal structure was published by Aller et al. [7] revealing up to now not known insights: P-gp is comprised of two transmembrane domains (TMDs) and two nucleotide binding domains (NBDs) spanning ~136Å perpendicular to and ~70Å in the plane of the bilayer. The distance between the NBDs averages ~30Å. The NBD is situated in the cytoplasm and responsible for binding and hydrolyzing ATP to provide the energy for the efflux process. The TMD consists of two bundles of six alpha helices reaching throughout the membrane bilayer: TMs 1 to 3, 6, 10, 11 and TMs 4, 5, 7 to 9, 12. P-gp binds a wide range of substrates in this region and changes its conformation to pump substances out of the cell. The binding pocket is mostly formed by hydrophobic and aromatic residues. It offers a lot of space (internal cavity within the lipid bilayer is ~6000Å^3) as it is six times bigger than that of BmrR (transcription regulator from Bacillus subtilis) accommodating inter alia lipids, sterols, peptides and metabolic products [7] [**Figure 5**].

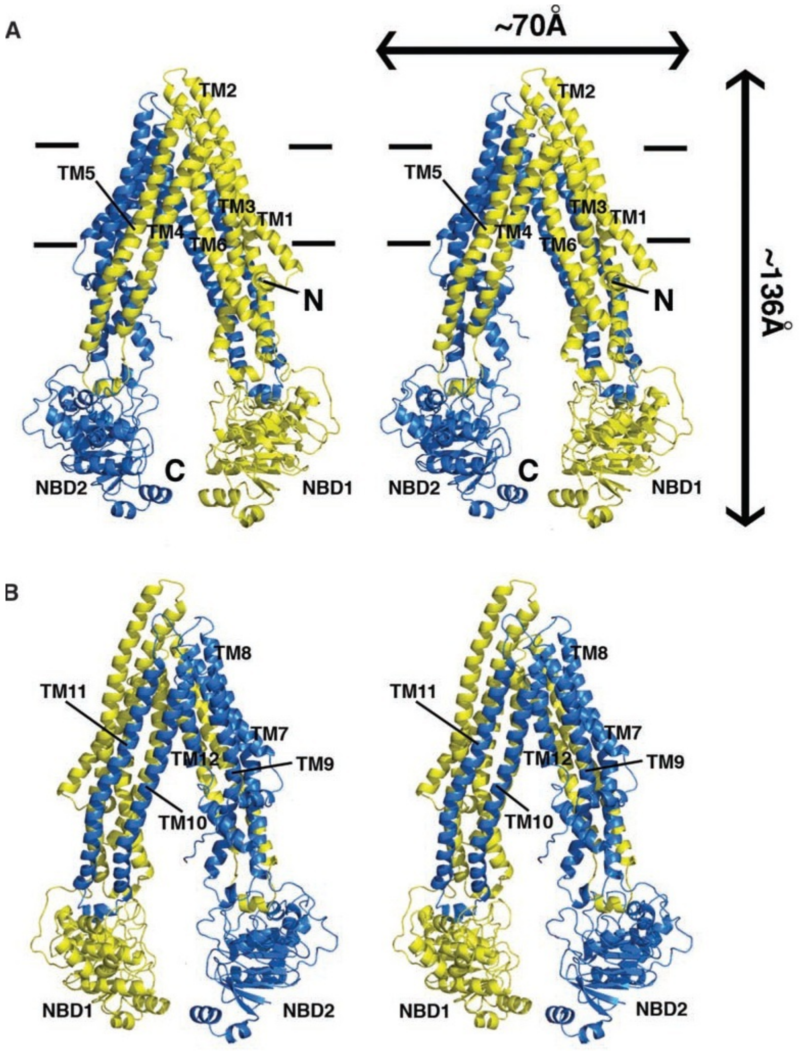


Figure 5 *– Front and back view of P-gp, Figure taken from [7].*

Genes encoding the P-gp are divided into 2 classes in humans (MDR1 and MDR3/MDR2) and 3 members in mice (mdr3/mdr1a, mdr1/mdr1b and mdr2) and rats (pgp1, pgp2/mdr1b and pgp3) [8] as shown **in Table 2**.

Table 2 *- Classification of P-gp isoforms [8]*

Species	*Class I*	*Class II*	*Class III*
Human	*MDR1*		*MDR3/MDR2*
Mouse	*mdr3/mdr1a*	*mdr1/mdr1b*	*mdr2*
Rat	*pgp1*	*pgp2/mdr1b*	*pgp3*

The sequence conservation of the P-gp gene family across species is very high. Class III is more than 90% identical between mice, hamsters and humans. In humans classes I and III are 75% identical. In mice the highest levels of mdr1 (class II) are described in pregnant uterus, adrenals, placenta and kidney, while mdr2 (class III) is mostly expressed in the liver and muscle and mdr3 (class I) is frequently detected in the intestine and lung. Moreover, the profile of mdr gene expression is conserved across species: human MDR1 (class I) expression is overlapping with that of mouse mdr1 and mdr3 and human MDR3 (class III) expression is overlapping with mouse mdr2. In rats mdr2 (class III) is highly expressed in the liver, muscle, heart and spleen and at lower levels in the lung and brain, whereas mdr1b (class II) is frequently detected in the lung and rarely in the liver, kidney, small intestine and spleen[8].

As mentioned before P-gp is a member of the MDR subfamily and therefore plays a role in multidrug resistance. The protein encoded by the MDR gene effluxes xenobiotic compounds with broad substrate specificity and as a result decreases drug accumulation in multidrug-resistant cells. The over expression of P-gp is one reason for the resistance of tumor cells to multiple chemotherapeutic drugs [**Figure 6**].

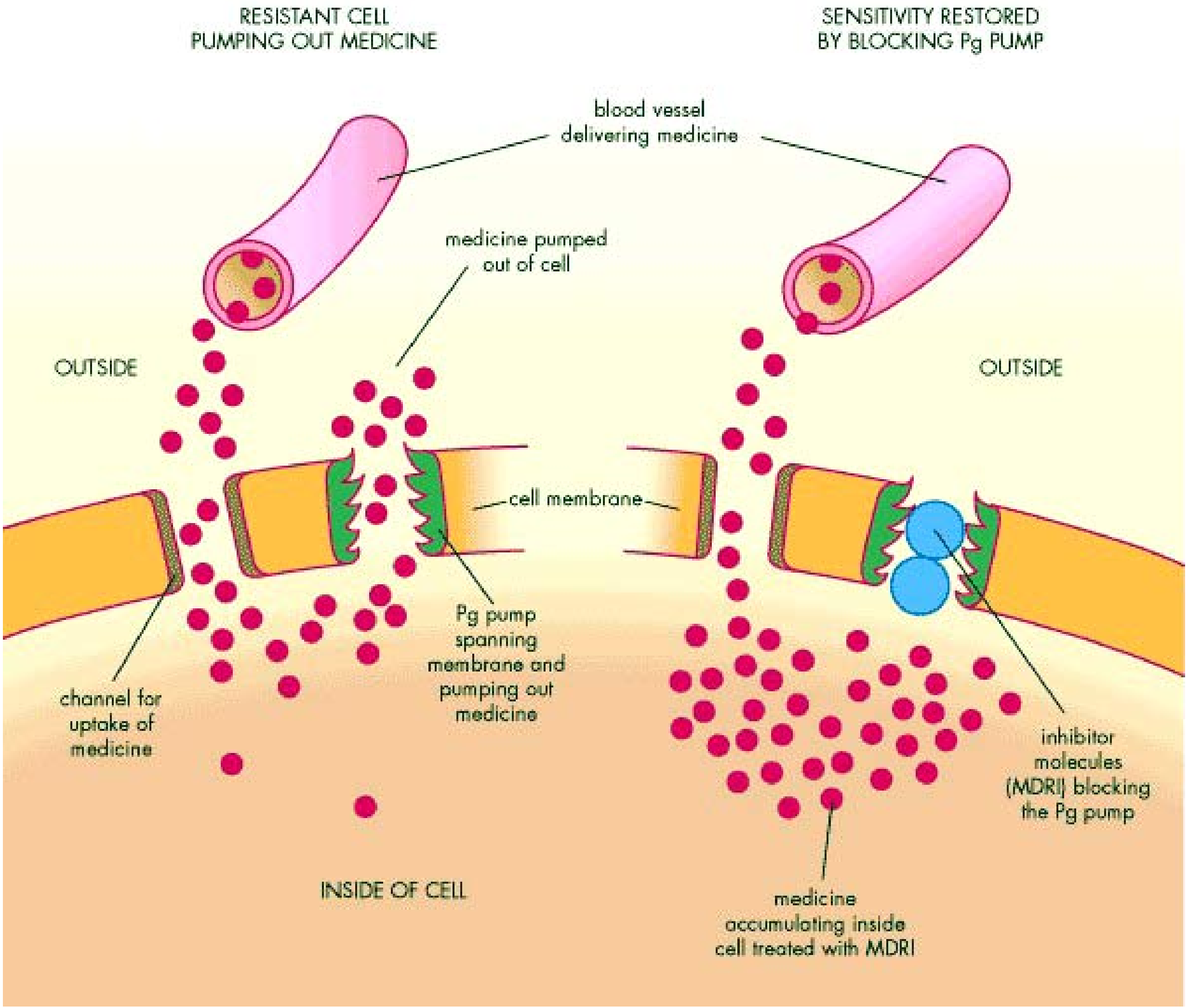

RESISTANT CELL
PUMPING OUT MEDICINE
SENSITIVITY RESTORED
BY BLOCKING Pg PUMP
blood vessel
delivering medicine
medicine pumped
out of cell
OUTSIDE
OUTSIDE
cell membrane
Pg pump
spanning
membrane and
pumping out
medicine
channel for
uptake of
medicine
inhibitor
molecules
(MDRI) blocking
the Pg pump
INSIDE OF CELL
medicine
accumulating inside
cell treated with MDRI

Figure 6 *– An illustration of how multidrug resistance inhibitors (MDRIs) can block the P-gp of resistant tumor cells, Figure taken from [9].*

C. Computational Background

1. Molecular Modeling

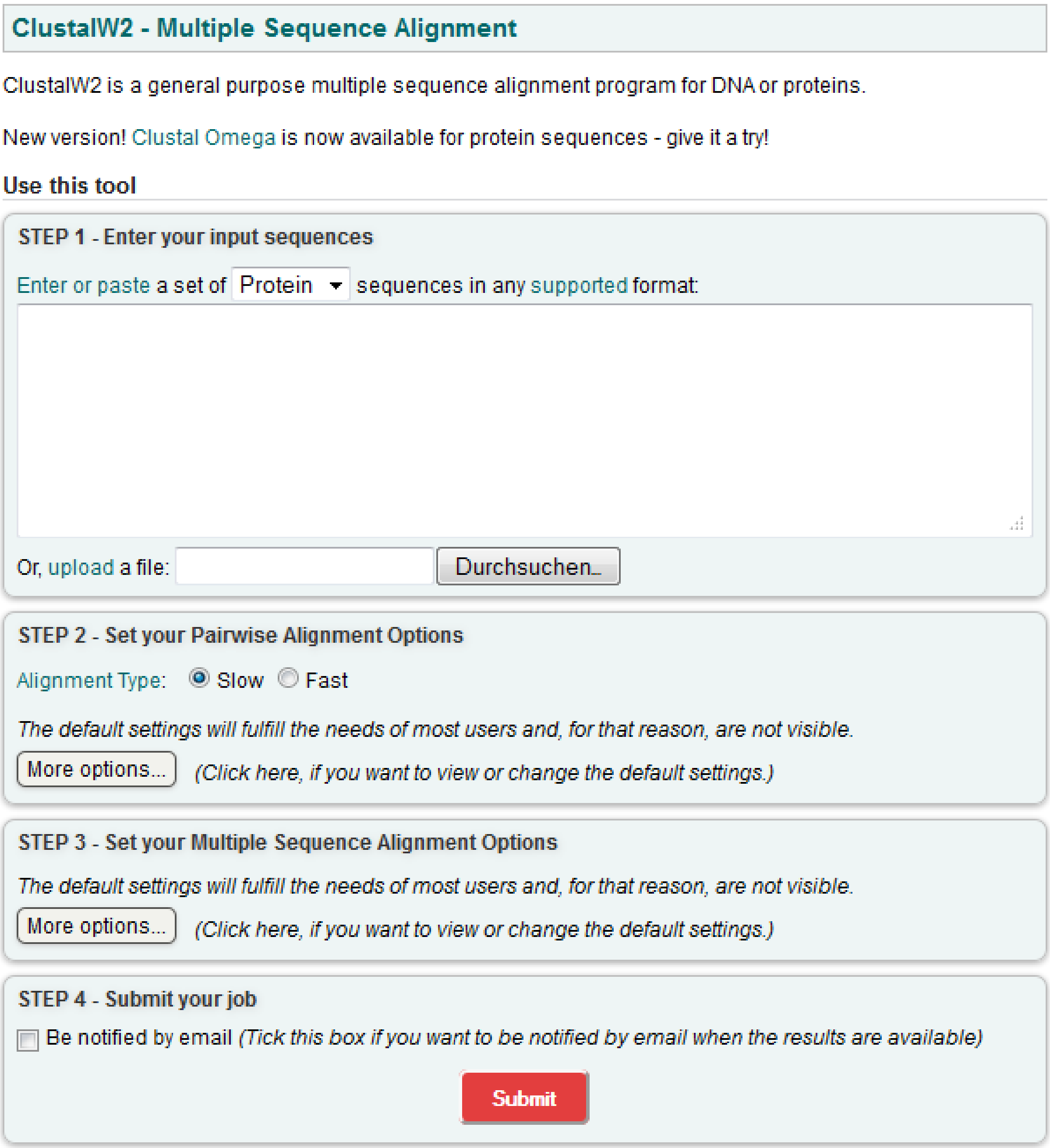
All theoretical methods and computational techniques used to model or mimic the behavior of molecules are covered by molecular modeling. The techniques are exploited in the fields of computational chemistry, computational biology and materials science for exploring molecular systems ranging from small chemical systems to large biological molecules and material assemblies. The simplest calculations can be executed by hand, but unavoidably computers are required to perform molecular modeling of any reasonably sized system. The atomistic level description of the molecular systems can be seen as the common feature of molecular modeling techniques; individual atoms are the lowest level of information. The advantage of molecular modeling is the reduction of the complexity of the system that allows considering more atoms during calculations.

2. Sequence Alignment

In bioinformatics the sequence alignment is a most widely used tool to analyze DNA, RNA or protein similarity. It is routinely a part of more complicated analysis pipelines, like homology modeling (see page 25). Alignments are important for highlighting areas of similarity which may be associated with specific features that have been more highly conserved than other regions [10]. Two methods are known to carry out an alignment: pair wise sequence alignment and multiple sequence alignment. The pair wise sequence alignment is used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences. By contrast, the multiple sequence alignment aligns three or more biological sequences of similar length. Multiple sequence alignment is an important step for phylogenetic analysis, which intends to model the substitutions that have happened over evolution and obtain the evolutionary relationships between sequences. Several packages are available, e.g. ClustalW, ClustalX, T-Coffee, MAFFT and MUSCLE [11].

a. ClustalW

ClustalW is a tool to align three or more sequences together in a computationally efficient manner. ClustalW multiple sequence alignment is offered for free. The web form [**Figure 7**] is available at http://www.ebi.ac.uk/Tools/msa/clustalw2/.



ClustalW2 - Multiple Sequence Alignment

ClustalW2 is a general purpose multiple sequence alignment program for DNA or proteins.

New version! Clustal Omega is now available for protein sequences - give it a try!

Use this tool

STEP 1 - Enter your input sequences

Enter or paste a set of Protein sequences in any supported format:

Or, upload a file:

STEP 2 - Set your Pairwise Alignment Options

Alignment Type: ● Slow ○ Fast

The default settings will fulfill the needs of most users and, for that reason, are not visible.

(Click here, if you want to view or change the default settings.)

STEP 3 - Set your Multiple Sequence Alignment Options

The default settings will fulfill the needs of most users and, for that reason, are not visible.

(Click here, if you want to view or change the default settings.)

STEP 4 - Submit your job

☐ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

***Figure 7** – Depiction of ClustalW web input form, Figure taken from [10].*

There exist two ways to utilize the service at EBI: interactively or by e-Mail. The interactive way displays the results in the browser window. When the e-Mail option is chosen a link to the results will be sent by mail. The program accepts nucleic acid or protein sequences in the following multiple sequence input format:

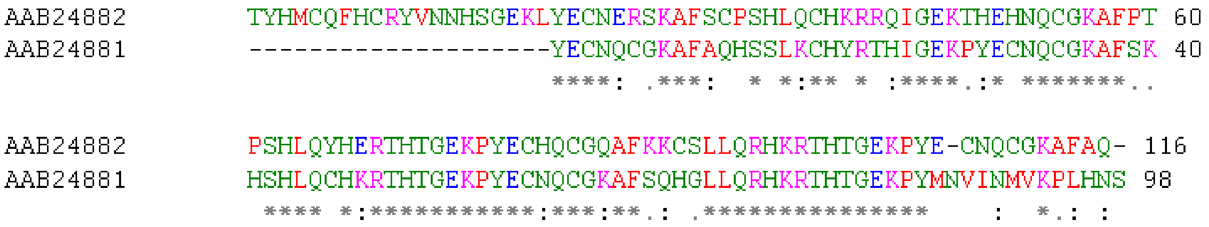
- NBRF/PIR
- EMBL/UniProt
- Pearson (FASTA)
- GDE

- ALN/ClustalW
- GCG/MSF
- RSF

For the alignment the sequences can either be pasted into the web form or uploaded to the web form in a file. It is very important that each of the sequences has a unique name. If they do not, the program will fail. Other reasons for failure are empty lines, white spaces or control characters between sequences or at the top of the file. The input for ClustalW is limited to a maximum of 500 sequences or to a 1MB file. When the input file or number of sequences is too large ClustalW can run for days and in some cases may not finish at all. For larger amounts of data the e-Mail results option should be used. The alignment method can be set to slow but accurate, or fast but approximate. ClustalW produces several outputs depending on the selected options when submitting the job. The output format for the alignment file can be as follows:

- ALN/ClustalW with base/residue numbering
- ALN/ClustalW without base/residue numbering
- GCG MSF
- PHYLIP
- NEXUS
- NBRF/PIR
- GDE
- Pearson/FASTA

By default the main output is the alignment file [**Figure 8**]. Other outputs can be downloaded in the results summary tab. The ClustalW output contains a Scores Table that shows the pair wise scores calculated for every pair of sequences that is to be aligned. Pair wise scores are the number of identities between the two sequences, divided by the length of the alignment, and represented as a percentage. This alignment is only a forerunner to the full multiple alignment [10].



```
AAB24882        TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT 60
AAB24881        --------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK 40
                                    ****: .***:  * *:** * :****.:* *******..

AAB24882        PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE-CNQCGKAFAQ- 116
AAB24881        HSHLQCHKRTHTGEKPYECNQCGKAFSQHGLLQRHKRTHTGEKPYMNVINMVKPLHNS 98
                 **** *:***********:***:**.: .***************    :  *.: :
```

*__Figure 8__ – A sequence alignment of two human zinc finger proteins, calculated by ClustalW and identified on the left by Gen Bank accession number. An * (asterisk) indicates positions which have a single, fully conserved residue. A : (colon) indicates conservation between groups of strongly similar properties (same color group). A . (period) indicates conservation between groups of weakly similar properties (similar shapes), edited from [10].*

The residue colors according to their physicochemical properties:

Residue	***Color***	***Property***
AVFPMILW	*RED*	*Small (small+ hydrophobic (incl.aromatic -Y))*
DE	*BLUE*	*Acidic*
RK	*MAGENTA*	*Basic - H*
STYHCNGQ	*GREEN*	*Hydroxyl + sulfhydryl + amine + G*
Others	*Grey*	*Unusual amino/imino acids etc*

3. Homology Modeling

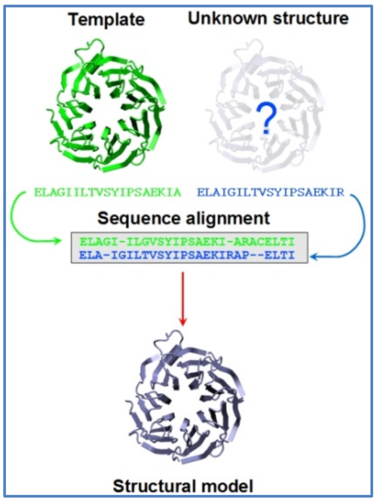

Template
Unknown structure
?
ELAGIILTVSYIPSAEKIA
ELAIGILTVSYIPSAEKIR
Sequence alignment
ELAGI-ILGVSYIPSAEKI-ARACELTI
ELA-IGILTVSYIPSAEKIRAP--ELTI
Structural model

***Figure 9** – Basic concept of Homology Modeling. For an unknown target structure with a known protein target sequence a homologous structural resolved protein is searched via sequence alignment. This protein is then served as a structural template for the target sequence, Figure taken from [12].*

With the techniques' development in molecular biology rapid identification, isolation and sequencing of genes became possible and enabled to infer the sequences of many proteins. A major goal of structural biology is the prediction of the three-dimensional structure from the sequence. Unfortunately this aim hasn't been reached until now. Nevertheless, alternative strategies allow developing models of protein structure when the X-ray or NMR structure is not available.

One method to calculate reasonable models of protein structures is homology modeling. This approach uses a "target" protein from its amino acid sequence and an experimental three-dimensional structure of a related homologous "template" protein for model building [**Figure 9**]. Homology modeling is based on the identification of one or more known protein structures resembling the structure of the query sequence and on the calculation of an alignment mapping residues in the query sequence to residues in the template sequence [**Figure 10**].

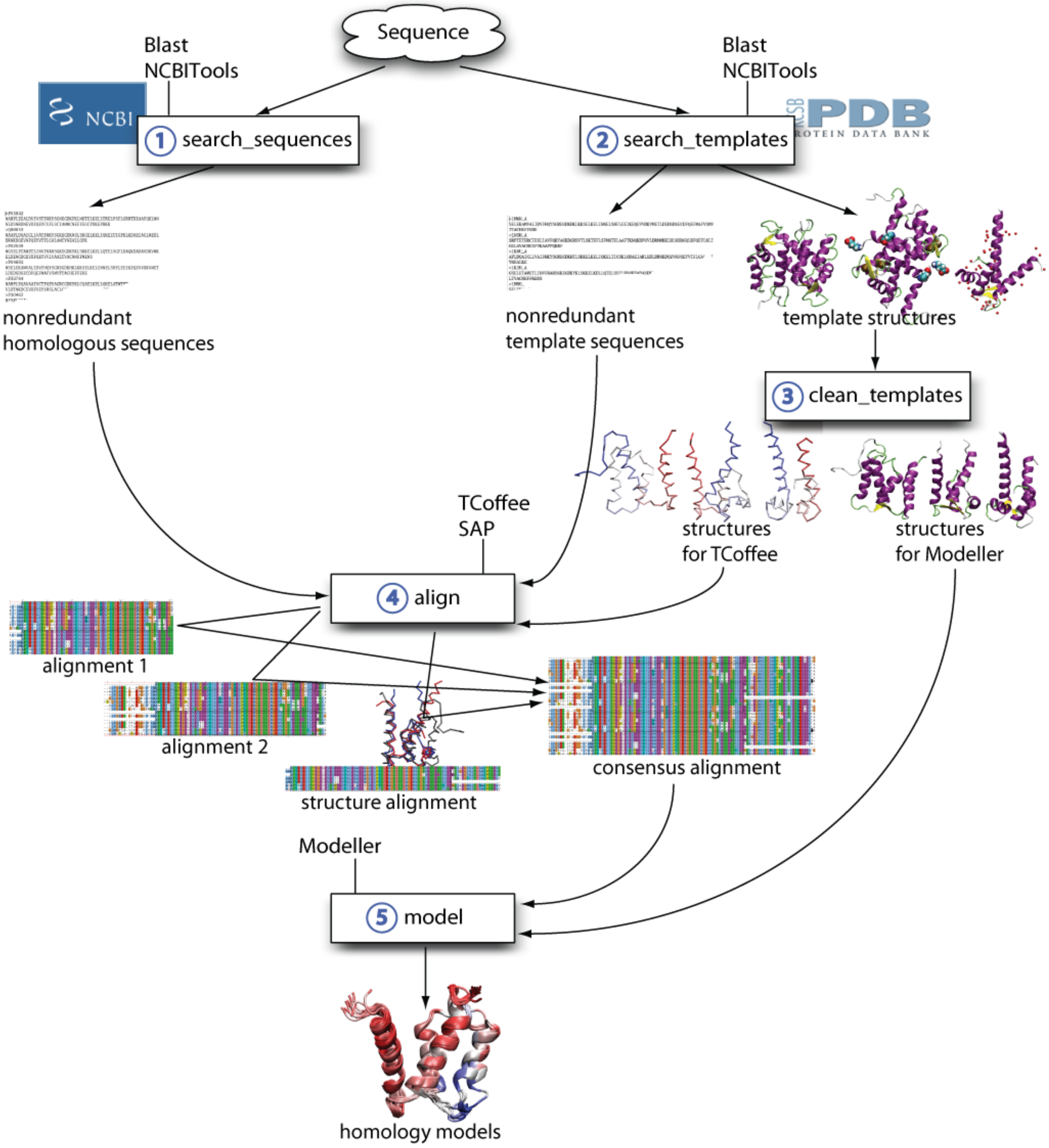

Sequence
Blast
NCBITools
NCBI
1 search_sequences
Blast
NCBITools
PDB
PROTEIN DATA BANK
2 search_templates
nonredundant
homologous sequences
nonredundant
template sequences
template structures
3 clean_templates
TCoffee
SAP
structures
for TCoffee
structures
for Modeller
4 align
alignment 1
alignment 2
structure alignment
consensus alignment
Modeller
5 model
homology models

Figure 10 *– Scheme of Homology modeling in more detail, Figure taken from [13].*

1. *Search for homologous sequences*
2. *Search for homologous sequences with known 3D structure*
3. *Cleaning the PDB files for the subsequent steps*
4. *Determine sequence alignments between target and templates*
5. *Finally building the structural models based on aligned sequence and structural template*

The sequence alignment and template structure are responsible for the quality of the homology model. Alignment gaps complicate the calculation and decrease the quality because they indicate that a structural region is present in the target but not in the template.

Accordingly model quality declines with decreasing sequence identity: high-accuracy comparative models are based on more than 50% sequence identity to their templates, medium-accuracy models on 30-50% identity and finally low-accuracy models on less than 30% sequence identity [14].

a. Modeller

Modeller is a computer program that calculates three-dimensional structures of proteins and their assemblies by satisfaction of spatial restraints. It is most commonly utilized for homology or comparative protein structure modeling. The program works with a scripting language and does not include any graphics. It will run on Windows, Mac or UNIX. For the calculation an alignment of an amino acid sequence that has to be modeled and a known related structure is needed [**Figure 11**].

```
>P1;3g5u_pajeva.pdb
structureX:3g5u_pajeva.pdb:   33:B : 1271 :B :::3.80:0.347
VSVLTMFRYAGWLDRLYML
VGTLAAIIHGVALPLMMLIFGDMTDSFASVGN--VSKNSTNMSEADKRAM
FAK--LEEEMTTYAYYYTGIGAGVLIVAYIQVSFWCLAAGRQIHKIRQKF
FHAIMNQEIGWFDVHDVGELNTRLTDDVSKINEGIGDKIGMFFQAMATFF
GGFIIGFTRGWKLTLVILAISPVLGLSAGIWAKILSSFTDKELHAYAKAG
AVAEEVLAAIRTVIAFGGQKKELERYNNNLEEAKRLGIKKAITANISMGA
AFLLIYASYALAFWYGTSLVISKEYSIGQVLTVFFSVLIGAFSVGQASPN
IEAFANARGAAYEVFKIIDNKPSIDSFSKSGHKPDNIQGNLEFKNIHFSY
PSRKEVQILKGLNLKVKSGQTVALVGNSGCGKSTTVQLMQRLYDPLDGMV
SIDGQDIRTINVRYLREIIGVVSQEPVLFATTIAENIRYGREDVTMDEIE
KAVKEANAYDFIMKLPHQFDTLVGERGAQLSGGQKQRIAIARALVRNPKI
LLLDEATSALDTESEAVVQAALDKAREGRTTIVIAHRLSTVRNADVIAGF
DGGVIVEQGNHDELMREKGIYFKLVMTQT
LDEDVPPASFWRILK
LNSTEWPYFVVGIFCAIINGGLQPAFSVIFSKVVGVFTNGGPPETQRQNS
NLFSLLFLILGIISFITFFLQGFTFGKAGEILTKRLRYMVFKSMLRQDVS
WFDDPKNTTGALTTRLANDAAQVKGATGSRLAVIFQNIANLGTGIIIS--
LIYGWQLTLLLLAIVPIIAIAGVVEMKMLSGQALKDKKELEGSGKIATEA
IENFRTVVSLTREQKFETMYAQSLQIPYRNAMKKAHVFGITFSFTQAMMY
FSYAAAFRFGAYLVTQQLMTFENVLLVFSAIVFGAMAVGQVSSFAPDYAK
ATVSASHIIRIIEKTPEIDSYSTQGLKPNMLEGNVQFSGVVFNYPTRPSI
PVLQGLSLEVKKGQTLALVGSSGCGKSTVVQLLERFYDPMAGSVFLDGKE
IKQLNVQWLRAQLGIVSQEPILFDCSIAENIAYGDNSRVVSYEEIVRAAK
EANIHQFIDSLPDKYNTRVGDKGTQLSGGQKQRIAIARALVRQPHILLLD
EATSALDTESEKVVQEALDKAREGRTCIVIAHRLSTIQNADLIVVIQNGK
VKEHGTHQQLLAQKGIYFSMVSVQA---
*
```

```
>P1;MDR1_RAT
sequence:MDR1_RAT:    : :    : ::::
VGIFGMFRYADWLDKLCMA
LGTLAAIIHGTLLPLLMLVFGYMTDSFTPSRDPHSDRAITNQSEINSTHT
VSDTSLEEDMAMYAYYYTGIGAGVLIVAYIQVSLWCLAAGRQIHKIRQKF
FHAIMNQEIGWFDVNDAGELNTRLTDDVSKINDGIGDKLGMFFQSITTFS
AGFIIGFISGWKLTLVILAVSPLIGLSSAMWAKVLTSFTNKELQAYAKAG
AVAEEVLAAIRTVIAFGGQKKELERYNKNLEEAKRVGIKKAITANISIGI
AYLLVYASYALAFWYGTSLVLSNEYSIGQVLTVFFSILLGTFSIGHLAPN
IEAFANARGAAYEIFKIIDNEPSIDSFSTKGHKPDSIMGNLEFKNVYFNY
PSRSEVKILKGLNLKVKSGQTVALVGNSGCGKSTTVQLLQRLYDPIEGEV
SIDGQDIRTINVRYLREIIGVVSQEPVLFATTIAENIRYGRENVTMDEIE
KAVKEANAYDFIMKLPHKFDTLVGERGAQLSGGQKQRIAIARALVRNPKI
LLLDEATSALDTESEAVVQAALDKAREGRTTIVIAHRLSTVRNADVIAGF
DGGVIVEQGNHEELMKEKGIYFKLVMTQT/
VDEDVPMVSFWQILK
LNISEWPYLVVGVLCAVINGCIQPVFAIVFSKIVGVFSRDDDHETKQRNC
NLFSLLFLVMGMISFVTYFFQGFTFGKAGEILTKRLRYMVFKSMLRQDIS
WFDDHKNTTGSLTTRLASDASNVKGAMGSRLAVVTQNVANLGTGIILSLV
LVYGWQLTLLLVVIIPLIVLGGIIEMKLLSGQALKDKKELEISGKIATEA
IENFRTVVSLTREQKFETMYAQSLQIPYRNALKKAHVFGITFAFTQAMIY
FSYAACFRFGAYLVARELMTFENVMLVFSAVVFGAMAAGNTSSFAPDYAK
AKVSASHIIGIIEKIPEIDSYSTEGLKPNWLEGNVKFNGVKFNYPTRPNI
PVLQGLSFEVKKGQTLRLVGSSGCGKSTVVQLLERFYNPMAGTVFLDGKE
IKQLNVQCVRA-LGIVSQEPILFDCSIAENIAYGDNSRVVSHEEIVRAAR
EANIHQFIDSLPEKYNTRVGDKGTQLSGGQKQRIAIARALVRQPHILLLD
EATSALDTESEKVVQEALDKAREGRTCVVIAHRLSTIQNADLIVVIQNGQ
VKEHGTHQQLLAQKGIYFSMVQAGAKRS
*
```

Figure 11 *– Depiction of an aligned sequence in pir format. The template sequence (PDB entry 3G5U) is given in the first part. The second part shows the alignment of the template sequence (MDR1_RAT). The * symbols sign the end of each sequence.*

With the command “mod9.10 model-default.py” Modeller automatically calculates a model with all non-hydrogen atoms.

“First many distance and dihedral angle restraints on the target sequence are calculated from its alignment with template 3D structures. The form of these restraints was obtained from a statistical analysis of the relationships between many pairs of homologous structures. This analysis relied on a database of 105 family alignments that included 416 proteins with known 3D structure. By scanning the database, tables quantifying various correlations were obtained, such as the correlations between two equivalent Cα-Cα distances or between equivalent main chain dihedral angles from two related proteins. These relationships were expressed as conditional probability density functions (pdf) and can be used directly as spatial restraints.

Next, the spatial restraints and CHARMM energy terms enforcing proper stereochemistry are combined into and objective function. Finally, the model is obtained by optimizing the objective function in Cartesian space. The optimization is carried out by the use of the variable target function method employing methods of conjugate gradients and molecular dynamics with simulated annealing. Several slightly different models can be calculated by varying the initial structure. The variability among these models can be used to estimate the errors in the corresponding regions of the fold [15] [**Figure 12**]."

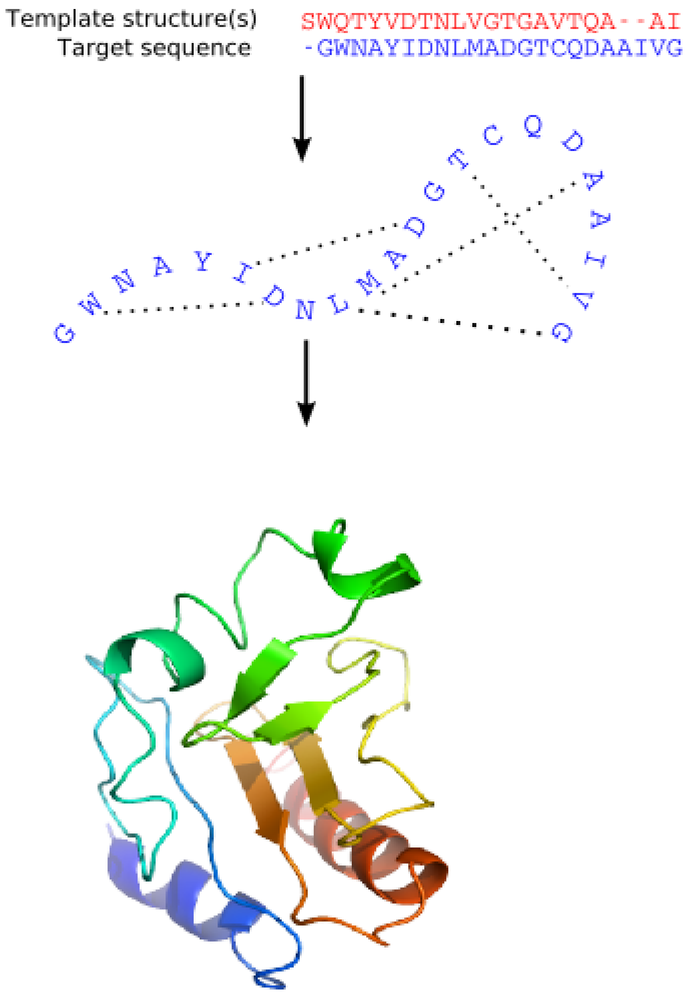


Figure 12 *– Scheme of building a homology model within MODELLER. 1. The known template 3D structure is aligned with the target sequence to be modeled. 2. Spatial features, as Cα-Cα distances, hydrogen bonds and main chain/side chain dihedral angles are extracted from the template and transferred to the target. 3. The 3D model is obtained by satisfying all the restraints as good as possible, taken from [15].*

For evaluation Modeller offers the molecular PDF (molpdf), which is the sum of all restraints, the GA-341 score, which assesses the overall fold quality and the "discrete optimized energy" score (DOPE). The molpdf and DOPE score are not absolute measures therefore they can only be utilized to rank models. Molpdf is specific for a set of restraints and DOPE for a target sequence. The molpdf and the DOPE score should be as low as possible and the GA341 score ranges from 0 (worst) to 1 (native like). However, the GA341 score is not as good as the DOPE score at distinguishing well from bad models [15].

Additionally, Modeller is able to perform multiple comparisons of protein sequences or structures, clustering of proteins and searching of sequence databases.

4. Docking

Docking calculations have been used in pharmaceutical research for nearly two decades. Virtual screening on protein templates, which varies from molecular similarity- and ligand-based virtual screening methods, offers an opportunity for the de novo identification of active compounds, without favoritism towards known hits or leads.

In the field of molecular modeling, docking appears as a computational simulation which foresees the preferred orientation of a molecule to a second one when bound to each other [16]. More precisely the docking process involves the prediction of ligand conformation and orientation within a targeted binding site and the prediction of the binding affinity [17].

Most of the computer studies on molecular docking assume one of the docking partners to be a protein, also called “receptor” or “receiving molecule”. On the other hand there is the complementary partner molecule which binds to the receptor, named “ligand”. During the first step posing samples the ligands’ translational, rotational and conformational degrees of freedom within the active site (see a, page 31). After this calculation, different poses or binding modes can be evaluated with the scoring function (see b, page 32), which counts the number of favorable intermolecular interactions such as hydrogen bonds and hydrophobic contacts. In the end the ranking classifies which ligands most likely interact favorably with a particular receptor based on the assigned scoring values.

The problem with molecular docking can be seen as a “lock and key” issue. In this case the protein is represented by the “lock” and the ligand by a “key”. During the docking calculation the protein and the ligand alter their conformation to achieve the “best fit” orientation, also known as “induced fit”.

a. Posing

In this initial step searching algorithms sample ligand orientations within the binding site. Most docking programs consider the protein as rigid and ligand flexibility is treated mainly by three categories [17]:

1. Simulation methods (molecular dynamics, energy minimization)
2. Random or stochastic methods (Monte Carlo, genetic algorithms, tabu search)
3. Systematic methods (incremental construction, conformational search, databases)

1. Simulation methods are implemented in the following software packages

- DOCK
- Glide
- MOE-Dock
- AutoDock
- Hammerhead

2. Random or stochastic methods are implemented in the following software packages

- AutoDock (MC)
- MOE-Dock (MC,TS)
- GOLD (GA)
- PRO_LEADS (TS)

3. Systematic methods are implemented in the following software packages

- DOCK (incremental)
- FlexX (incremental)
- Glide (incremental)
- Hammerhead (incremental)
- FLOG (database)

The handling of protein flexibility is less advanced than that of ligand flexibility, but various approaches have been utilized to flexibly calculate at least part of the target, including molecular dynamics and Monte Carlo calculations, rotamer libraries and protein ensemble grids [17].

b. Scoring

After the posing the fit complementarity of the generated ligand-receptor complexes is evaluated by a scoring function. This function attempts to estimate the binding free energy of the complex with computational algorithms which sums up calculated ligand-receptor interactions. Scoring functions implemented in docking programs make various assumptions and simplifications in the evaluation of modeled complexes and do not fully account for a number of physical phenomena that determine molecular recognition, e.g. entropic effects. Basically three different types of scoring functions can be distinguished [17]:

Types of scoring functions:

1. Force-field-based
    - D-Score
    - G-Score
    - GOLD
    - AutoDock
    - DOCK
2. Empirical
    - LUDI
    - F-Score
    - ChemScore
    - SCORE
    - Fresno
    - X-Score
3. Knowledge-based
    - PMF
    - DrugScore
    - SMoG

c. GOLD

GOLD is a program which calculates the docking modes of small molecules in protein binding sites. It is offered as a part of the program GOLD Suite, also containing Hermes for structure visualization and manipulation, and GOLDMine for post-processing of docking results. As mentioned on page 31 GOLD uses a genetic algorithm (GA) for protein-ligand docking. A GA is a computer program that imitates the process of evolution [18]. “It sets up a population of potential solutions at random. Each member of the population is encoded as a chromosome, which contains information about the mapping of protein ligand interactions. Every chromosome is assigned a fitness score based on its predicted binding affinity and the chromosomes within the population are ranked according to fitness [19]."

The ligand can be kept fully flexible, the protein partially flexible or it is possible to dock into multiple models of the same or different proteins. GOLD accepts mol2, mol and pdb as ligand input files and pdb and mol2 as protein file formats. The docking run can be launched with the help of the set-up wizard via Hermes. Before the calculation can be started the following preparations have to be done:

- GA speed settings
- loading the protein target
- specifying the binding site
- uploading the ligand(s)
- selecting the number of dockings and early termination allowance
- choosing the fitness function

GOLD provides three different scoring functions: GOLDScore, ChemScore and ASP. All of them calculate fitness scores that are dimensionless. The score illustrates how good the pose is; the higher the score, the better the docking result. The GOLDScore fitness function is the original scoring function offered with GOLD and is the one selected by default. It has been developed for the prediction of ligand binding positions and takes into consideration factors such as H-bonding energy, van der Waals energy and ligand torsion strain [19].

GOLD Fitness = $S_{hb\_ext} + S_{vdw\_ext} + S_{hb\_int} + S_{vdw\_int}$

S_{hb_ext}: ***protein-ligand hydrogen-bond score***

S_{vdw_ext}: ***protein-ligand van der Waals score***

S_{hb_int}: ***contribution to the Fitness due to intramolecular hydrogen bonds in the ligand***

S_{vdw_int}: ***contribution due to intramolecular strain in the ligand*** [20]

The ChemScore fitness function assesses the total free energy change that occurs on ligand binding and was trained by regression against binding affinity data. The ASP fitness function is an atom-atom potential obtained from a database of protein-ligand complexes and can be likened to other such scoring potentials, e.g. PMF and Drugscore. ASP integrates some

ChemScore terms. As the fitness scores are dimensionless they cannot be utilized explicitly as values for binding energy or binding affinity [19].

II. Aim of the study

In early stage drug development the pharmacokinetic profile and the possible toxicity of a drug candidate are determined in animal models (usually mouse or rat) before it is tested in human beings. Thus predicting toxicity only in humans during the clinical trials is far too late. First pharmacokinetic and toxicological tests are carried out in animals several years before the drug candidate is even admitted for testing in humans. Thus, besides developing predictive in silico models for the identification of ligands for human P-gp it is also important to establish predictive models for mouse and rat P-gp. Furthermore, early in silico prediction of in vivo toxicological outcomes might increase the quality of drug candidates, lower the attrition rate during subsequent phases of drug development, and reduce the number of animals to be used in preclinical studies.

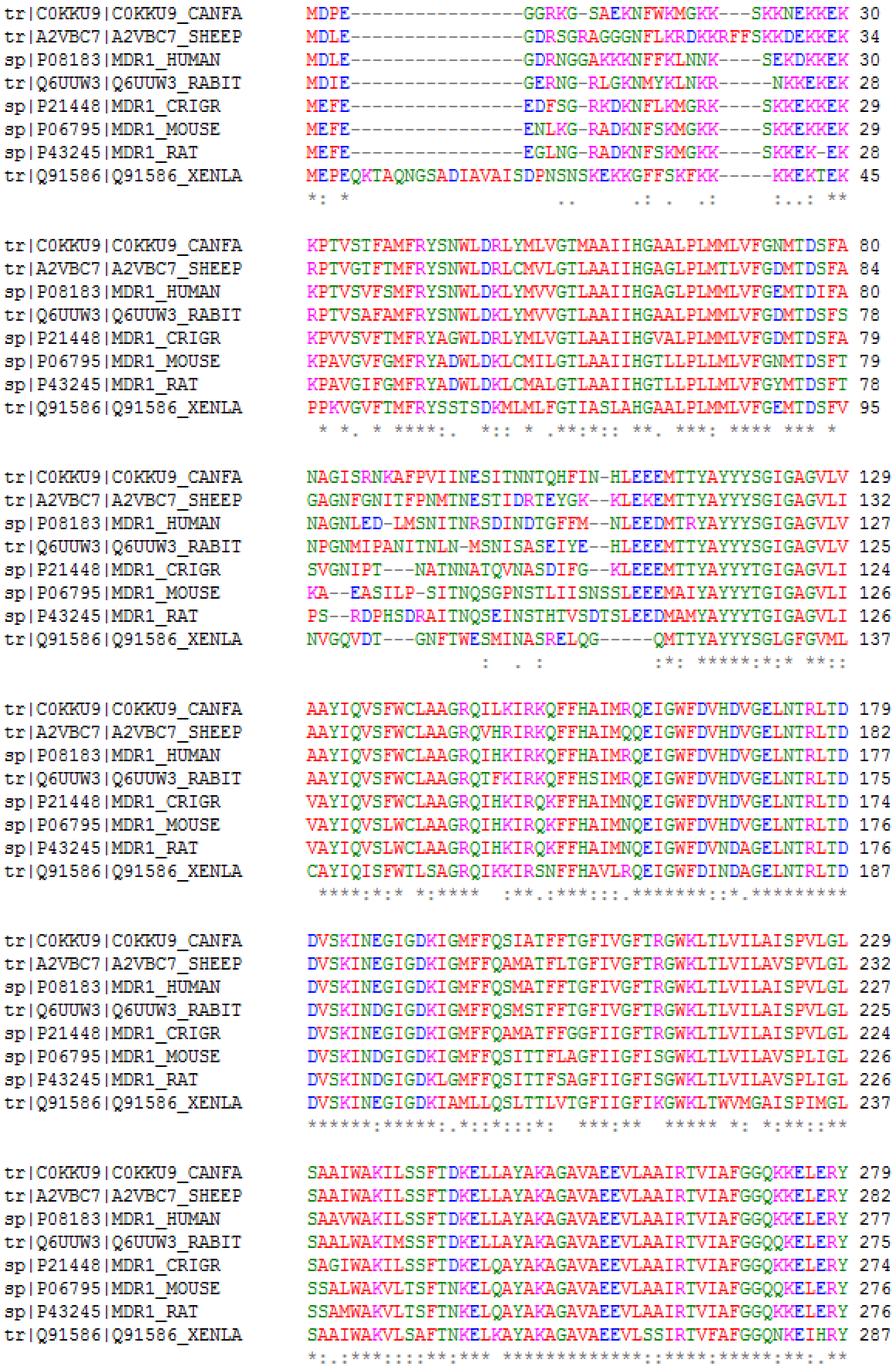
The difficulty in structure-based in silico studies with membrane proteins like P-gp is the fact that due to technical difficulties in the crystallizing process, X-ray structures and other high-resolution structural data are mostly unavailable. Therefore, computational methods such as homology modeling and docking are needed to explore molecular binding modes. However, in case of P-gp, since 2009 the mouse crystal structure is available [7]. As the rat P-gp shares high sequence identity (92%) to the recent crystallized mouse P-gp, we used it as a template for a rat P-gp homology model.

The obtained protein homology model will be validated using routine methods. Subsequently, the model will be used for docking of known rat P-gp inhibitors into the rat P-gp homology model. The resulting docking ranking list will then be compared to the known IC_{50} values of these already published and tested inhibitors. Hence this comparison will be used to evaluate the predictive potency of the model.

III. Materials and Methods

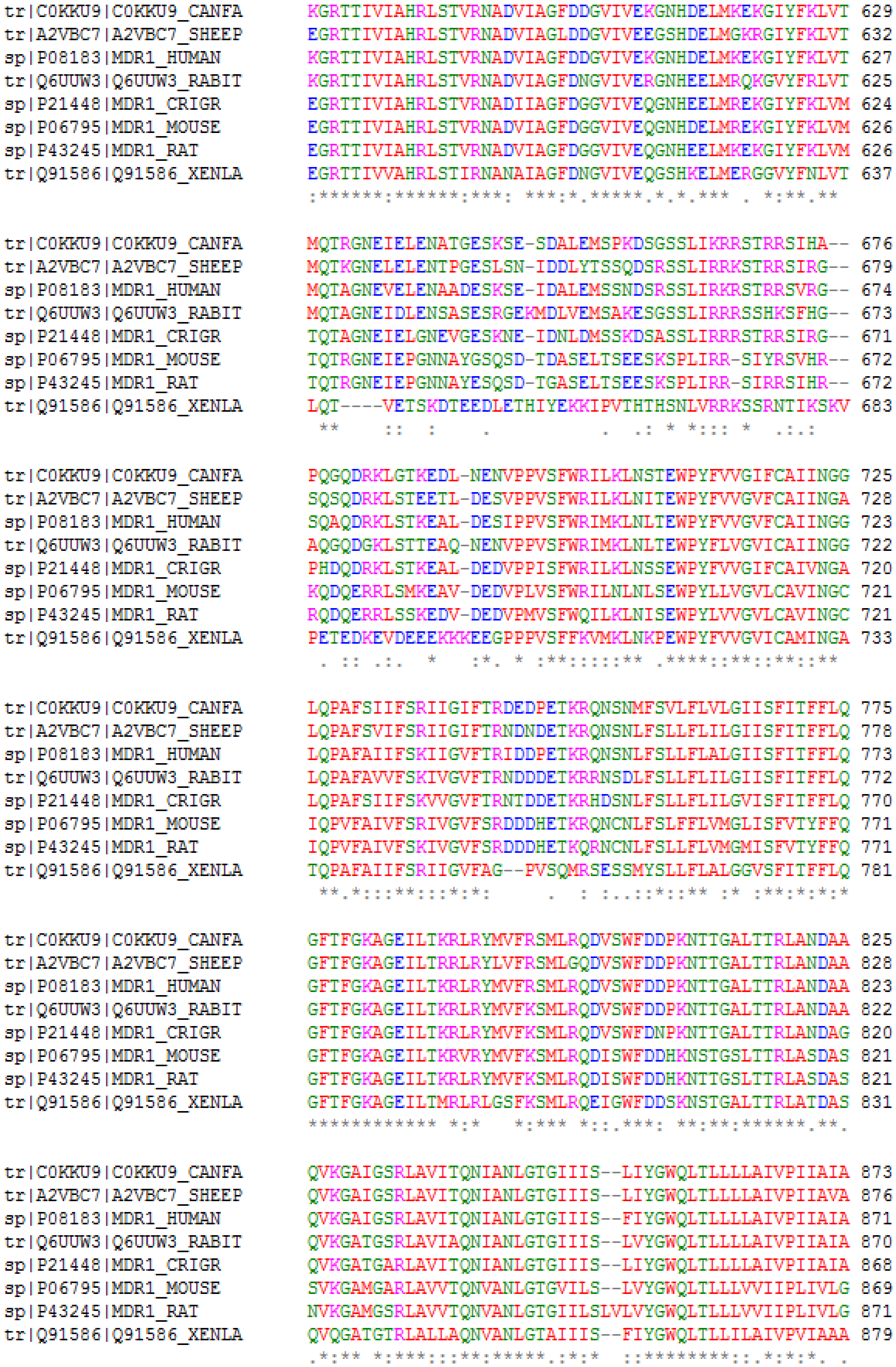
A. Multiple sequence alignment

The P-gp sequences of different species (dog, frog, hamster, human, mouse, rabbit, rat and sheep) were compared with ClustalW version 2.1 [21, 22] taking the whole protein as well as the transmembrane domains only. All settings were used as default. **Figure 13** shows the alignment of the whole P-gp sequences.



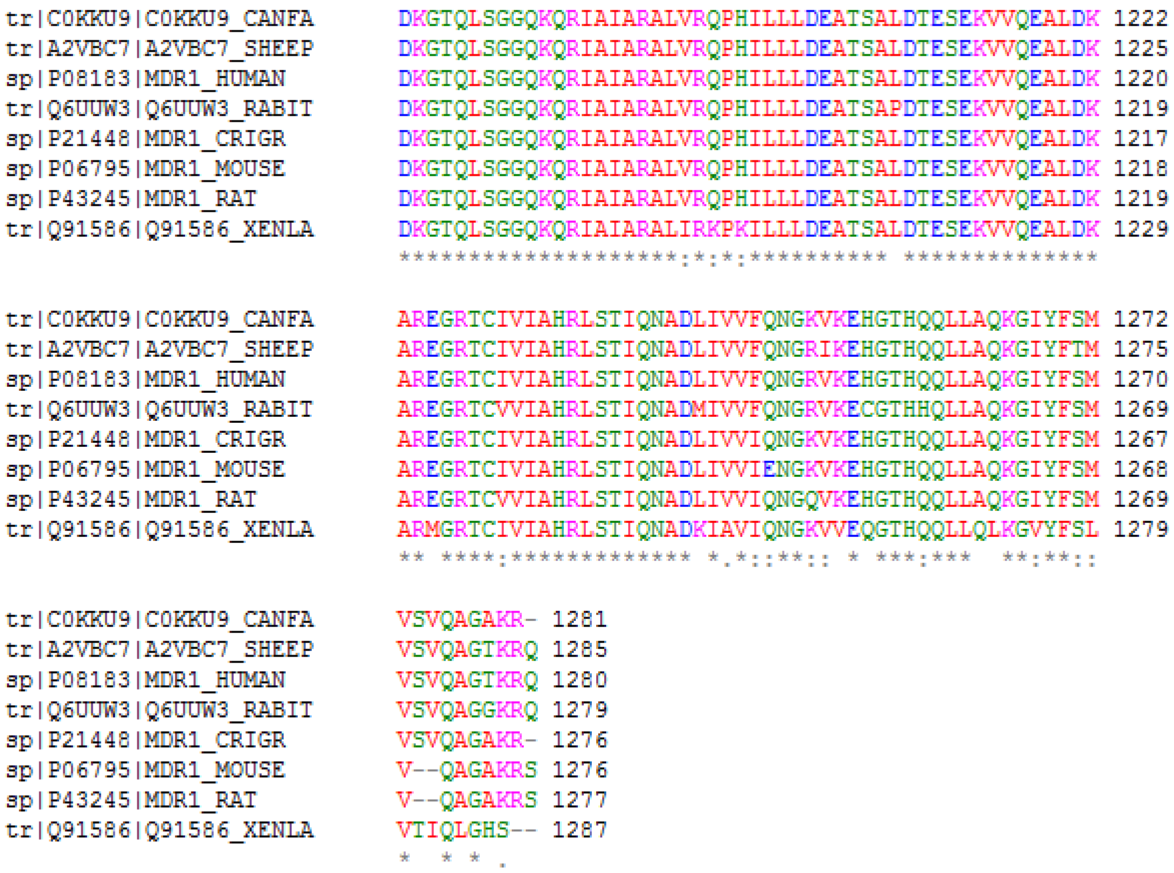
```
tr|C0KKU9|C0KKU9_CANFA          MDPE----------------GGRKG-SAEKNFWKMGKK---SKKNEKKEK 30
tr|A2VBC7|A2VBC7_SHEEP          MDLE----------------GDRSGRAGGGNFLKRDKKRFFSKKDEKKEK 34
sp|P08183|MDR1_HUMAN            MDLE----------------GDRNGGAKKKNFFKLNNK----SEKDKKEK 30
tr|Q6UUW3|Q6UUW3_RABIT          MDIE----------------GERNG-RLGKNMYKLNKR-----NKKEKEK 28
sp|P21448|MDR1_CRIGR            MEFE----------------EDFSG-RKDKNFLKMGRK----SKKEKKEK 29
sp|P06795|MDR1_MOUSE            MEFE----------------ENLKG-RADKNFSKMGKK----SKKEKKEK 29
sp|P43245|MDR1_RAT              MEFE----------------EGLNG-RADKNFSKMGKK----SKKEK-EK 28
tr|Q91586|Q91586_XENLA          MEPEQKTAQNGSADIAVAISDPNSNSKEKKGFFSKFKK-----KKEKTEK 45
                                *: *                   ..     .: .  .:     :..: **

tr|C0KKU9|C0KKU9_CANFA          KPTVSTFAMFRYSNWLDRLYMLVGTMAAIIHGAALPLMMLVFGNMTDSFA 80
tr|A2VBC7|A2VBC7_SHEEP          RPTVGTFTMFRYSNWLDRLCMVLGTLAAIIHGAGLPLMTLVFGDMTDSFA 84
sp|P08183|MDR1_HUMAN            KPTVSVFSMFRYSNWLDKLYMVVGTLAAIIHGAGLPLMMLVFGEMTDIFA 80
tr|Q6UUW3|Q6UUW3_RABIT          RPTVSAFAMFRYSNWLDKLYMVVGTLAAIIHGAALPLMMLVFGDMTDSFS 78
sp|P21448|MDR1_CRIGR            KPVVSVFTMFRYAGWLDRLYMLVGTLAAIIHGVALPLMMLVFGDMTDSFA 79
sp|P06795|MDR1_MOUSE            KPAVGVFGMFRYADWLDKLCMILGTLAAIIHGTLLPLLMLVFGNMTDSFT 79
sp|P43245|MDR1_RAT              KPAVGIFGMFRYADWLDKLCMALGTLAAIIHGTLLPLLMLVFGYMTDSFT 78
tr|Q91586|Q91586_XENLA          PPKVGVFTMFRYSSTSDKMLMLFGTIASLAHGAALPLMMLVFGEMTDSFV 95
                                 * *. * ****:.  *:: * .**:*:: **. ***: **** *** *

tr|C0KKU9|C0KKU9_CANFA          NAGISRNKAFPVIINESITNNTQHFIN-HLEEEMTTYAYYYSGIGAGVLV 129
tr|A2VBC7|A2VBC7_SHEEP          GAGNFGNITFPNMTNESTIDRTEYGK--KLEKEMTTYAYYYSGIGAGVLI 132
sp|P08183|MDR1_HUMAN            NAGNLED-LMSNITNRSDINDTGFFM--NLEEDMTRYAYYYSGIGAGVLV 127
tr|Q6UUW3|Q6UUW3_RABIT          NPGNMIPANITNLN-MSNISASEIYE--HLEEEMTTYAYYYSGIGAGVLV 125
sp|P21448|MDR1_CRIGR            SVGNIPT---NATNNATQVNASDIFG--KLEEEMTTYAYYYTGIGAGVLI 124
sp|P06795|MDR1_MOUSE            KA--EASILP-SITNQSGPNSTLIISNSSLEEEMAIYAYYYTGIGAGVLI 126
sp|P43245|MDR1_RAT              PS--RDPHSDRAITNQSEINSTHTVSDTSLEEDMAMYAYYYTGIGAGVLI 126
tr|Q91586|Q91586_XENLA          NVGQVDT---GNFTWESMINASRELQG-----QMTTYAYYYSGLGFGVML 137
                                                :  . :          :*: *****:*:* **::

tr|C0KKU9|C0KKU9_CANFA          AAYIQVSFWCLAAGRQILKIRKQFFHAIMRQEIGWFDVHDVGELNTRLTD 179
tr|A2VBC7|A2VBC7_SHEEP          AAYIQVSFWCLAAGRQVHRIRKQFFHAIMQQEIGWFDVHDVGELNTRLTD 182
sp|P08183|MDR1_HUMAN            AAYIQVSFWCLAAGRQIHKIRKQFFHAIMRQEIGWFDVHDVGELNTRLTD 177
tr|Q6UUW3|Q6UUW3_RABIT          AAYIQVSFWCLAAGRQTFKIRKQFFHSIMRQEIGWFDVHDVGELNTRLTD 175
sp|P21448|MDR1_CRIGR            VAYIQVSFWCLAAGRQIHKIRQKFFHAIMNQEIGWFDVHDVGELNTRLTD 174
sp|P06795|MDR1_MOUSE            VAYIQVSLWCLAAGRQIHKIRQKFFHAIMNQEIGWFDVHDVGELNTRLTD 176
sp|P43245|MDR1_RAT              VAYIQVSLWCLAAGRQIHKIRQKFFHAIMNQEIGWFDVNDAGELNTRLTD 176
tr|Q91586|Q91586_XENLA          CAYIQISFWTLSAGRQIKKIRSNFFHAVLRQEIGWFDINDAGELNTRLTD 187
                                 ****:*:* *:****  :**.:***:::.*******::*.*********

tr|C0KKU9|C0KKU9_CANFA          DVSKINEGIGDKIGMFFQSIATFFTGFIVGFTRGWKLTLVILAISPVLGL 229
tr|A2VBC7|A2VBC7_SHEEP          DVSKINEGIGDKIGMFFQAMATFLTGFIVGFTRGWKLTLVILAVSPVLGL 232
sp|P08183|MDR1_HUMAN            DVSKINEGIGDKIGMFFQSMATFFTGFIVGFTRGWKLTLVILAISPVLGL 227
tr|Q6UUW3|Q6UUW3_RABIT          DVSKINDGIGDKIGMFFQSMSTFFTGFIVGFTRGWKLTLVILAISPVLGL 225
sp|P21448|MDR1_CRIGR            DVSKINEGIGDKIGMFFQAMATFFGGFIIGFTRGWKLTLVILAISPVLGL 224
sp|P06795|MDR1_MOUSE            DVSKINDGIGDKIGMFFQSITTFLAGFIIGFISGWKLTLVILAVSPLIGL 226
sp|P43245|MDR1_RAT              DVSKINDGIGDKLGMFFQSITTFSAGFIIGFISGWKLTLVILAVSPLIGL 226
tr|Q91586|Q91586_XENLA          DVSKINEGIGDKIAMLLQSLTTLVTGFIIGFIKGWKLTWVMGAISPIMGL 237
                                ******:*****:.*::*:::*:  ***:**  ***** *: *:**::**

tr|C0KKU9|C0KKU9_CANFA          SAAIWAKILSSFTDKELLAYAKAGAVAEEVLAAIRTVIAFGGQKKELERY 279
tr|A2VBC7|A2VBC7_SHEEP          SAAIWAKILSSFTDKELLAYAKAGAVAEEVLAAIRTVIAFGGQKKELERY 282
sp|P08183|MDR1_HUMAN            SAAVWAKILSSFTDKELLAYAKAGAVAEEVLAAIRTVIAFGGQKKELERY 277
tr|Q6UUW3|Q6UUW3_RABIT          SAALWAKIMSSFTDKELLAYAKAGAVAEEVLAAIRTVIAFGGQQKELERY 275
sp|P21448|MDR1_CRIGR            SAGIWAKILSSFTDKELQAYAKAGAVAEEVLAAIRTVIAFGGQKKELERY 274
sp|P06795|MDR1_MOUSE            SSALWAKVLTSFTNKELQAYAKAGAVAEEVLAAIRTVIAFGGQQKELERY 276
sp|P43245|MDR1_RAT              SSAMWAKVLTSFTNKELQAYAKAGAVAEEVLAAIRTVIAFGGQKKELERY 276
tr|Q91586|Q91586_XENLA          SAAIWAKVLSAFTNKELKAYAKAGAVAEEVLSSIRTVFAFGGQNKEIHRY 287
                                *:.:***::::**:*** *************::****:*****:**:.**
```

```
tr|C0KKU9|C0KKU9_CANFA      NKNLEEAKRIGIKKAITANISIGAAFLLIYASYALAFWYGTSLVLSSEYS 329
tr|A2VBC7|A2VBC7_SHEEP      NKNLEEAKRIGIKKAITANISMGAAFLLMYASYALAFWYGTSLVLSREYS 332
sp|P08183|MDR1_HUMAN        NKNLEEAKRIGIKKAITANISIGAAFLLIYASYALAFWYGTTLVLSGEYS 327
tr|Q6UUW3|Q6UUW3_RABIT      NKNLEEAKRIGIKKAITANISVGVAFLLMYASYALAFWYWNHLGHLKEYS 325
sp|P21448|MDR1_CRIGR        NNNLEEAKRLGIKKAITANISMGAAFLLIYASYALAFWYGTSLVISKEYS 324
sp|P06795|MDR1_MOUSE        NKNLEEAKNVGIKKAITASISIGIAYLLVYASYALAFWYGTSLVLSNEYS 326
sp|P43245|MDR1_RAT          NKNLEEAKRVGIKKAITANISIGIAYLLVYASYALAFWYGTSLVLSNEYS 326
tr|Q91586|Q91586_XENLA      EKNLEDAKKIGIKKAITANVSIGFAFLMIYAAYSLAFWYGTTLIIDGGYT 337
                            ::***:**.:********.:*:* *:*::**:*:***** . *     *:

tr|C0KKU9|C0KKU9_CANFA      IGQVLTVFFSVLIGAFSIGQASPSIEAFANARGAAYEIFKIIDNKPSIDS 379
tr|A2VBC7|A2VBC7_SHEEP      IGQVLTVFFSVLLGTFSIGQASPNIEAFANARGAAYEVFKIIDNKPSINS 382
sp|P08183|MDR1_HUMAN        IGQVLTVFFSVLIGAFSVGQASPSIEAFANARGAAYEIFKIIDNKPSIDS 377
tr|Q6UUW3|Q6UUW3_RABIT      IGQVLTVFFSVLVGAFSIGQASPNVEAFANARGAAYEIFRIIDNMPSIDS 375
sp|P21448|MDR1_CRIGR        IGQVLTVFFAVLIGAFSIGQASPNIEAFANARGAAYEIFNIIDNKPSIDS 374
sp|P06795|MDR1_MOUSE        IGEVLTVFFSILLGTFSIGHLAPNIEAFANARGAAFEIFKIIDNEPSIDS 376
sp|P43245|MDR1_RAT          IGQVLTVFFSILLGTFSIGHLAPNIEAFANARGAAYEIFKIIDNEPSIDS 376
tr|Q91586|Q91586_XENLA      IGSVLTVFFAVIIGAFAVGQTSPNIEAFANARGAAYTIFNIIDNQPKIDS 387
                            **.******::::*:*::*: :*.:**********: :*.**** *.*:*

tr|C0KKU9|C0KKU9_CANFA      YSKSGHKPDNIKGNLEFKNVHFSYPSRKEVKILKGLNLKVQSGQTVALVG 429
tr|A2VBC7|A2VBC7_SHEEP      YSNAGHKPDNIKGNLEFRNVHFHYPSRNEVKILKGLNLKVGSGQTVALVG 432
sp|P08183|MDR1_HUMAN        YSKSGHKPDNIKGNLEFRNVHFSYPSRKEVKILKGLNLKVQSGQTVALVG 427
tr|Q6UUW3|Q6UUW3_RABIT      YSEAGHKPDNIKGNLEFRNVHFSYPSRKEVKILKGLNLKVESGQTVALVG 425
sp|P21448|MDR1_CRIGR        FSKNGYKPDNIKGNLEFKNIHFSYPSRKDVQILKGLNLKVQSGQTVALVG 424
sp|P06795|MDR1_MOUSE        FSTKGYKPDSIMGNLEFKNVHFNYPSRSEVQILKGLNLKVKSGQTVALVG 426
sp|P43245|MDR1_RAT          FSTKGHKPDSIMGNLEFKNVYFNYPSRSEVKILKGLNLKVKSGQTVALVG 426
tr|Q91586|Q91586_XENLA      FSKEGLKPDKIKGDIEFKNVIFTYPSRKDIQVLKGLNLNIPSGKTVALVG 437
                            :*  * ***.* *::**:*: * ****.::::******:: **:******

tr|C0KKU9|C0KKU9_CANFA      NSGCGKSTTVQLMQRLYDPTDGMVCIDGQDIRTINVRHLREITGVVSQEP 479
tr|A2VBC7|A2VBC7_SHEEP      NSGCGKSTTVQLMQRLYDPTEGMVSIDGQDIRTINVRYLREIIGVVSQEP 482
sp|P08183|MDR1_HUMAN        NSGCGKSTTVQLMQRLYDPTEGMVSVDGQDIRTINVRFLREIIGVVSQEP 477
tr|Q6UUW3|Q6UUW3_RABIT      NSGCGKSTTVQLMRRLYDPTDGVVSIDGQDIRTMNVRYLREITGVVSQEP 475
sp|P21448|MDR1_CRIGR        NSGCGKSTTVQLLQRLYDPTEGVVSIDGQDIRTINVRYLREIIGVVSQEP 474
sp|P06795|MDR1_MOUSE        NSGCGKSTTVQLMQRLYDPLEGVVSIDGQDIRTINVRYLREIIGVVSQEP 476
sp|P43245|MDR1_RAT          NSGCGKSTTVQLLQRLYDPIEGEVSIDGQDIRTINVRYLREIIGVVSQEP 476
tr|Q91586|Q91586_XENLA      SSGCGKSTTVQLIQRFYDPEDGVITLDGQDIRSLNIRYLREIIGVVSQEP 487
                            .***********::*:*** :* : :******::*:*.**** *******

tr|C0KKU9|C0KKU9_CANFA      VLFATTIAENIRYGRENVTMDEIEKAVKEANAYDFIMKLPNKFDTLVGER 529
tr|A2VBC7|A2VBC7_SHEEP      VLFATTIAENIRYGREDVTMDEIQKAVKEANAYDFIMKLPNKFDTLVGER 532
sp|P08183|MDR1_HUMAN        VLFATTIAENIRYGRENVTMDEIEKAVKEANAYDFIMKLPHKFDTLVGER 527
tr|Q6UUW3|Q6UUW3_RABIT      VLFATTIAENVRYGREDVTMDEIEKAVKEANAYNFIMKLPHKFDTLVGER 525
sp|P21448|MDR1_CRIGR        VLFATTIAENIRYGRENVTMDEIEKAVKEANAYDFIMKLPHKFDTLVGER 524
sp|P06795|MDR1_MOUSE        VLFATTIAENIRYGREDVTMDEIEKAVKEANAYDFIMKLPHQFDTLVGER 526
sp|P43245|MDR1_RAT          VLFATTIAENIRYGRENVTMDEIEKAVKEANAYDFIMKLPHKFDTLVGER 526
tr|Q91586|Q91586_XENLA      ILFDTTIADNIRYGREDVTKEEIERATKEANAYDFIMKLPDKLETLVGER 537
                            :** ****:*:*****:** :**::*.******:******.:::******

tr|C0KKU9|C0KKU9_CANFA      GAQLSGGQKQRIAIARALVRNPKILLLDEATSALDTESEAVVQVALDKAR 579
tr|A2VBC7|A2VBC7_SHEEP      GAQLSGGQKQRIAIARALVRNPKILLLDEATSALDTESEAVVQAALDKAR 582
sp|P08183|MDR1_HUMAN        GAQLSGGQKQRIAIARALVRNPKILLLDEATSALDTESEAVVQVALDKAR 577
tr|Q6UUW3|Q6UUW3_RABIT      GAQLSGGQKQRIAIARALVRNPKILLLDEATSALDTESEAVVQVALDKAR 575
sp|P21448|MDR1_CRIGR        GAQLSGGQKQRIAIARALVRNPKILLLDEATSALDTESEAVVQAALDKAR 574
sp|P06795|MDR1_MOUSE        GAQLSGGQKQRIAIARALVRNPKILLLDEATSALDTESEAVVQAALDKAR 576
sp|P43245|MDR1_RAT          GAQLSGGQKQRIAIARALVRNPKILLLDEATSALDTESEAVVQAALDKAR 576
tr|Q91586|Q91586_XENLA      GTQLSGGQKQRIAIARALVRNPKILLLDEATSALDTESEAVVQSALDKAR 587
                            *:***************************************** ******
```



```
tr|C0KKU9|C0KKU9_CANFA        KGRTTIVIAHRLSTVRNADVIAGFDDGVIVEKGNHDELMKEKGIYFKLVT 629
tr|A2VBC7|A2VBC7_SHEEP        EGRTTIVIAHRLSTVRNADVIAGLDDGVIVEEGSHDELMGKRGIYFKLVT 632
sp|P08183|MDR1_HUMAN          KGRTTIVIAHRLSTVRNADVIAGFDDGVIVEKGNHDELMKEKGIYFKLVT 627
tr|Q6UUW3|Q6UUW3_RABIT        KGRTTIVIAHRLSTVRNADVIAGFDNGVIVERGNHEELMRQKGVYFRLVT 625
sp|P21448|MDR1_CRIGR          EGRTTIVIAHRLSTVRNADIIAGFDGGVIVEQGNHEELMREKGIYFKLVM 624
sp|P06795|MDR1_MOUSE          EGRTTIVIAHRLSTVRNADVIAGFDGGVIVEQGNHDELMREKGIYFKLVM 626
sp|P43245|MDR1_RAT            EGRTTIVIAHRLSTVRNADVIAGFDGGVIVEQGNHEELMKEKGIYFKLVM 626
tr|Q91586|Q91586_XENLA        EGRTTIVVAHRLSTIRNANAIAGFDNGVIVEQGSHKELMERGGVYFNLVT 637
                              :******:******:***: ***:*.*****.*.*.*** . *:**.**

tr|C0KKU9|C0KKU9_CANFA        MQTRGNEIELENATGESKSE-SDALEMSPKDSGSSLIKRRSTRRSIHA-- 676
tr|A2VBC7|A2VBC7_SHEEP        MQTKGNELELENTPGESLSN-IDDLYTSSQDSRSSLIRRKSTRRSIRG-- 679
sp|P08183|MDR1_HUMAN          MQTAGNEVELENAADESKSE-IDALEMSSNDSRSSLIRKRSTRRSVRG-- 674
tr|Q6UUW3|Q6UUW3_RABIT        MQTAGNEIDLENSASESRGEKMDLVEMSAKESGSSLIRRRSSHKSFHG-- 673
sp|P21448|MDR1_CRIGR          TQTAGNEIELGNEVGESKNE-IDNLDMSSKDSASSLIRRRSTRRSIRG-- 671
sp|P06795|MDR1_MOUSE          TQTRGNEIEPGNNAYGSQSD-TDASELTSEESKSPLIRR-SIYRSVHR-- 672
sp|P43245|MDR1_RAT            TQTRGNEIEPGNNAYESQSD-TGASELTSEESKSPLIRR-SIRRSIHR-- 672
tr|Q91586|Q91586_XENLA        LQT----VETSKDTEEDLETHIYEKKIPVTHTHSNLVRRKSSRNTIKSKV 683
                               **    ::  :    .          .  .: * *::: *  .:.:

tr|C0KKU9|C0KKU9_CANFA        PQGQDRKLGTKEDL-NENVPPVSFWRILKLNSTEWPYFVVGIFCAIINGG 725
tr|A2VBC7|A2VBC7_SHEEP        SQSQDRKLSTEETL-DESVPPVSFWRILKLNITEWPYFVVGVFCAIINGA 728
sp|P08183|MDR1_HUMAN          SQAQDRKLSTKEAL-DESIPPVSFWRIMKLNLTEWPYFVVGVFCAIINGG 723
tr|Q6UUW3|Q6UUW3_RABIT        AQGQDGKLSTTEAQ-NENVPPVSFWRIMKLNLTEWPYFLVGVICAIINGG 722
sp|P21448|MDR1_CRIGR          PHDQDRKLSTKEAL-DEDVPPISFWRILKLNSSEWPYFVVGIFCAIVNGA 720
sp|P06795|MDR1_MOUSE          KQDQERRLSMKEAV-DEDVPLVSFWRILNLNLSEWPYLLVGVLCAVINGC 721
sp|P43245|MDR1_RAT            RQDQERRLSSKEDV-DEDVPMVSFWQILKLNISEWPYLVVGVLCAVINGC 721
tr|Q91586|Q91586_XENLA        PETEDKEVDEEEKKKEEGPPPVSFFKVMKLNKPEWPYFVVGVICAMINGA 733
                               . :: .:.  *   :*. * :**:::::** .****::**::**::**

tr|C0KKU9|C0KKU9_CANFA        LQPAFSIIFSRIIGIFTRDEDPETKRQNSNMFSVLFLVLGIISFITFFLQ 775
tr|A2VBC7|A2VBC7_SHEEP        LQPAFSVIFSRIIGIFTRNDNDETKRQNSNLFSLLFLILGIISFITFFLQ 778
sp|P08183|MDR1_HUMAN          LQPAFAIIFSKIIGVFTRIDDPETKRQNSNLFSLLFLALGIISFITFFLQ 773
tr|Q6UUW3|Q6UUW3_RABIT        LQPAFAVVFSKIVGVFTRNDDDETKRRNSDLFSLLFLILGIISFITFFLQ 772
sp|P21448|MDR1_CRIGR          LQPAFSIIFSKVVGVFTRNTDDETKRHDSNLFSLLFLILGVISFITFFLQ 770
sp|P06795|MDR1_MOUSE          IQPVFAIVFSRIVGVFSRDDDHETKRQNCNLFSLFFLVMGLISFVTYFFQ 771
sp|P43245|MDR1_RAT            IQPVFAIVFSKIVGVFSRDDDHETKQRNCNLFSLLFLVMGMISFVTYFFQ 771
tr|Q91586|Q91586_XENLA        TQPAFAIIFSRIIGVFAG--PVSQMRSESSMYSLLFLALGGVSFITFFLQ 781
                               **.*:::**:::*:*:     .  : :..::*::** :* :**:*:*:*

tr|C0KKU9|C0KKU9_CANFA        GFTFGKAGEILTKRLRYMVFRSMLRQDVSWFDDPKNTTGALTTRLANDAA 825
tr|A2VBC7|A2VBC7_SHEEP        GFTFGKAGEILTRRLRYLVFRSMLGQDVSWFDDPKNTTGALTTRLANDAA 828
sp|P08183|MDR1_HUMAN          GFTFGKAGEILTKRLRYMVFRSMLRQDVSWFDDPKNTTGALTTRLANDAA 823
tr|Q6UUW3|Q6UUW3_RABIT        GFTFGKAGEILTKRLRYMVFKSMLRQDVSWFDDPKNTTGALTTRLANDAA 822
sp|P21448|MDR1_CRIGR          GFTFGKAGEILTKRLRYMVFKSMLRQDVSWFDNPKNTTGALTTRLANDAG 820
sp|P06795|MDR1_MOUSE          GFTFGKAGEILTKRVRYMVFKSMLRQDISWFDDHKNSTGSLTTRLASDAS 821
sp|P43245|MDR1_RAT            GFTFGKAGEILTKRLRYMVFKSMLRQDISWFDDHKNTTGSLTTRLASDAS 821
tr|Q91586|Q91586_XENLA        GFTFGKAGEILTMRLRLGSFKSMLRQEIGWFDDSKNSTGALTTRLATDAS 831
                              ************ *:*   *:*** *::.***: **:**:******.**.

tr|C0KKU9|C0KKU9_CANFA        QVKGAIGSRLAVITQNIANLGTGIIIS--LIYGWQLTLLLLAIVPIIAIA 873
tr|A2VBC7|A2VBC7_SHEEP        QVKGAIGSRLAVITQNIANLGTGIIIS--LIYGWQLTLLLLAIVPIIAVA 876
sp|P08183|MDR1_HUMAN          QVKGAIGSRLAVITQNIANLGTGIIIS--FIYGWQLTLLLLAIVPIIAIA 871
tr|Q6UUW3|Q6UUW3_RABIT        QVKGATGSRLAVIAQNIANLGTGIIIS--LVYGWQLTLLLLAIVPIIAIA 870
sp|P21448|MDR1_CRIGR          QVKGATGARLAVITQNIANLGTGIIIS--LIYGWQLTLLLLAIVPIIAIA 868
sp|P06795|MDR1_MOUSE          SVKGAMGARLAVVTQNVANLGTGVILS--LVYGWQLTLLLVVIIPLIVLG 869
sp|P43245|MDR1_RAT            NVKGAMGSRLAVVTQNVANLGTGIILSLVLVYGWQLTLLLVVIIPLIVLG 871
tr|Q91586|Q91586_XENLA        QVQGATGTRLALLAQNVANLGTAIIIS--FIYGWQLTLLILAIVPVIAAA 879
                              .*:** *:***:::**:*****.:*:*  ::********::.*:*:*. .
```

```
tr|C0KKU9|C0KKU9_CANFA      GVVEMKMLSGQALKDKKELEGAGKIATEAIENFRTVVSLTREQKFEYMYA 923
tr|A2VBC7|A2VBC7_SHEEP      GVIEMKMLSGQALKDKKELEGAGKIATEAIENFRTVVSLTREERFEYMYA 926
sp|P08183|MDR1_HUMAN        GVVEMKMLSGQALKDKKELEGSGKIATEAIENFRTVVSLTQEQKFEHMYA 921
tr|Q6UUW3|Q6UUW3_RABIT      GVVEMKMLSGQALKDKKELEGSGKIATEAIENFRTVVSLTREQKFENMYA 920
sp|P21448|MDR1_CRIGR        GVVEMKMLSGQALKDKKELEGSGKIATEAIENFRTVVSLTREQKFENMYA 918
sp|P06795|MDR1_MOUSE        GIIEMKLLSGQALKDKKQLEISGKIATEAIENFRTIVSLTREQKFETMYA 919
sp|P43245|MDR1_RAT          GIIEMKLLSGQALKDKKELEISGKIATEAIENFRTVVSLTREQKFETMYA 921
tr|Q91586|Q91586_XENLA      GLVEMKMFAGHAKKDKKELEKAGKISTDAVLNIRTVVSLTRERKFEAMYE 929
                            *::***:::*:* ****:** :***:*:*: *:**:****:*.:** **

tr|C0KKU9|C0KKU9_CANFA      QSLQVPYRNSLRKAHIFGVSFSITQAMMYFSYAGCFR-FGAYLVANEFMN 972
tr|A2VBC7|A2VBC7_SHEEP      QSLQVPYRNSLRKAHVFGITFSITQAMMYFSYAGCFR-FGAYLVAQGIME 975
sp|P08183|MDR1_HUMAN        QSLQVPYRNSLRKAHIFGITFSFTQAMMYFSYAGCFR-FGAYLVAHKLMS 970
tr|Q6UUW3|Q6UUW3_RABIT      QSLQVPYRNSLEKAHIFGITFSFTQAMMYFSYAGCFR-FGAFLVARELMS 969
sp|P21448|MDR1_CRIGR        QSLQIPYRNALKKAHVFGITFSFTQAMMYFSYAACFR-FGAYLVARELMT 967
sp|P06795|MDR1_MOUSE        QSLQVPYRNAMKKAHVFGITFSFTQAMMYFSYAACFR-FGAYLVAQQLMT 968
sp|P43245|MDR1_RAT          QSLQIPYRNALKKAHVFGITFAFTQAMIYFSYAACFR-FGAYLVARELMT 970
tr|Q91586|Q91586_XENLA      KSLEGPYRNSIKKAHLHGLTYGLSQAHHVLCLCWVFSVLGAYLVVEGLMK 979
                            :**: ****::.***:.*:::.::**   :. .  *  :**:**.. :*

tr|C0KKU9|C0KKU9_CANFA      FQDVLLVFSAIVFGAMAVGQVSSFAPDYAKAKVSAAHVIMIIEKSPLIDS 1022
tr|A2VBC7|A2VBC7_SHEEP      FQDVLLVFSAVVFGAMAVGQVSSFAPDYAKAKVSAAHVINIIEKIPLIDS 1025
sp|P08183|MDR1_HUMAN        FEDVLLVFSAVVFGAMAVGQVSSFAPDYAKAKISAAHIIMIIEKTPLIDS 1020
tr|Q6UUW3|Q6UUW3_RABIT      FENVLLVFSAVVFGAMAVGQVSSFAPDYAKAKISASHIIMILEKLPKIDS 1019
sp|P21448|MDR1_CRIGR        FENVLLVFSAIVFGAMAVGQVSSFAPDYAKAKVSASHIIMIIEKVPSIDS 1017
sp|P06795|MDR1_MOUSE        FENVMLVFSAVVFGAMAAGNTSSFAPDYAKAKVSASHIIRIIEKTPEIDS 1018
sp|P43245|MDR1_RAT          FENVMLVFSAVVFGAMAAGNTSSFAPDYAKAKVSASHIIGIIEKIPEIDS 1020
tr|Q91586|Q91586_XENLA      LDEVFLVSSAIVLGAMALGQTSSFAPDYTKAMISAAHIFSLLERVPQIDS 1029
                            :::*:** **:*:**** *:.*******:** :**:*:: ::*: * ***

tr|C0KKU9|C0KKU9_CANFA      YSPHGLKPNTLEGNVTFNEVVFNYPTRPDIPVLQGLSLEVKKGQTLALVG 1072
tr|A2VBC7|A2VBC7_SHEEP      YSTEGLKPSTVEGSVAFNDVVFNYPTRPDVPVLRGLSLEVKKGQTLALVG 1075
sp|P08183|MDR1_HUMAN        YSTEGLMPNTLEGNVTFGEVVFNYPTRPDIPVLQGLSLEVKKGQTLALVG 1070
tr|Q6UUW3|Q6UUW3_RABIT      YSTEGLKPGTLEGNMTFKDVVFNYPTRPDIPVLQGLNLQVKKGQTLALVG 1069
sp|P21448|MDR1_CRIGR        YSTGGLKPNTLEGNVKFNEVVFNYPTRPDIPVLQGLNLEVKKGQTLALVG 1067
sp|P06795|MDR1_MOUSE        YSTEGLKPTLLEGNVKFNGVQFNYPTRPNIPVLQGLSLEVKKGQTLALVG 1068
sp|P43245|MDR1_RAT          YSTEGLKPNWLEGNVKFNGVKFNYPTRPNIPVLQGLSFEVKKGQTLRLVG 1070
tr|Q91586|Q91586_XENLA      YSDQGEKPKNCSGNVVFKGVNFNYPTRPDITVLQGLDISVKQGETLALVG 1079
                            **  *  *   .*.: *  * *******::.**:**.:.**:*:** ***

tr|C0KKU9|C0KKU9_CANFA      SSGCGKSTVVQLLERFYDPLAGSVLIDGKEIKHLNVQWLRAHLGIVSQEP 1122
tr|A2VBC7|A2VBC7_SHEEP      SSGCGKSTVVQLLERFYDPLAGTVFIDGKEVKQLNVQWLRAHMGIVSQEP 1125
sp|P08183|MDR1_HUMAN        SSGCGKSTVVQLLERFYDPLAGKVLLDGKEIKRLNVQWLRAHLGIVSQEP 1120
tr|Q6UUW3|Q6UUW3_RABIT      PSGCGKSTVVQLIERFYDPLAGTVLLDGKEVNQLNVQWLRAHLGIVSQEP 1119
sp|P21448|MDR1_CRIGR        SSGCGKSTVVQLLERFYDPMAGTVFLDGKEVNQLNVQWLRAHLGIVSQEP 1117
sp|P06795|MDR1_MOUSE        SSGCGKSTVVQLLERFYDPMAGSVFLDGKEIKQLNVQWLRAHLGIVSQEP 1118
sp|P43245|MDR1_RAT          SSGCGKSTVVQLLERFYNPMAGTVFLDGKEIKQLNVQCVRA-LGIVSQEP 1119
tr|Q91586|Q91586_XENLA      SSGCGKSTTVSLLERFYDPFEGEVLVDGLSVRNLNIQWVRAQMGIVSQEP 1129
                            .*******.*.*:****:*: * *::** .:..**:* :** :*******

tr|C0KKU9|C0KKU9_CANFA      ILFDCSIAENIAYGDNSRVVSHEEIMQAAKEANIHHFIETLPEKYNTRVG 1172
tr|A2VBC7|A2VBC7_SHEEP      ILFDCSIGENIAYGDNSRVVSQEEIEHAAKEANIHSFIEMLPDKYNTRVG 1175
sp|P08183|MDR1_HUMAN        ILFDCSIAENIAYGDNSRVVSQEEIVRAAKEANIHAFIESLPNKYSTKVG 1170
tr|Q6UUW3|Q6UUW3_RABIT      ILFDCSIAENIAYGDNSRVVSQDEIIKAAKEANIHAFIDSLPDKYNTRVG 1169
sp|P21448|MDR1_CRIGR        ILFDCSIAENIAYGDNSRVVSQDEIERAAKEANIHQFIESLPDKYNTRVG 1167
sp|P06795|MDR1_MOUSE        ILFDCSIAENIAYGDNSRAVSHEEIVRAAKEANIHQFIDSLPDKYNTRVG 1168
sp|P43245|MDR1_RAT          ILFDCSIAENIAYGDNSRVVSHEEIVRAAREANIHQFIDSLPEKYNTRVG 1169
tr|Q91586|Q91586_XENLA      ILFDCSIGDNIAYGDNNRKVTQEEIETAAKEANIHSFIESLTDKYNTRVG 1179
                            *******.:*******.* *:::**  **:***** **: *.:**.*:**
```



```
tr|C0KKU9|C0KKU9_CANFA      DKGTQLSGGQKQRIAIARALVRQPHILLLDEATSALDTESEKVVQEALDK 1222
tr|A2VBC7|A2VBC7_SHEEP      DKGTQLSGGQKQRIAIARALVRQPHILLLDEATSALDTESEKVVQEALDK 1225
sp|P08183|MDR1_HUMAN        DKGTQLSGGQKQRIAIARALVRQPHILLLDEATSALDTESEKVVQEALDK 1220
tr|Q6UUW3|Q6UUW3_RABIT      DKGTQLSGGQKQRIAIARALVRQPHILLLDEATSAPDTESEKVVQEALDK 1219
sp|P21448|MDR1_CRIGR        DKGTQLSGGQKQRIAIARALVRQPHILLLDEATSALDTESEKVVQEALDK 1217
sp|P06795|MDR1_MOUSE        DKGTQLSGGQKQRIAIARALVRQPHILLLDEATSALDTESEKVVQEALDK 1218
sp|P43245|MDR1_RAT          DKGTQLSGGQKQRIAIARALVRQPHILLLDEATSALDTESEKVVQEALDK 1219
tr|Q91586|Q91586_XENLA      DKGTQLSGGQKQRIAIARALIRKPKILLLDEATSALDTESEKVVQEALDK 1229
                            ********************:*:*:********** **************

tr|C0KKU9|C0KKU9_CANFA      AREGRTCIVIAHRLSTIQNADLIVVFQNGKVKEHGTHQQLLAQKGIYFSM 1272
tr|A2VBC7|A2VBC7_SHEEP      AREGRTCIVIAHRLSTIQNADLIVVFQNGRIKEHGTHQQLLAQKGIYFTM 1275
sp|P08183|MDR1_HUMAN        AREGRTCIVIAHRLSTIQNADLIVVFQNGRVKEHGTHQQLLAQKGIYFSM 1270
tr|Q6UUW3|Q6UUW3_RABIT      AREGRTCVVIAHRLSTIQNADMIVVFQNGRVKECGTHHQLLAQKGIYFSM 1269
sp|P21448|MDR1_CRIGR        AREGRTCIVIAHRLSTIQNADLIVVIQNGKVKEHGTHQQLLAQKGIYFSM 1267
sp|P06795|MDR1_MOUSE        AREGRTCIVIAHRLSTIQNADLIVVIENGKVKEHGTHQQLLAQKGIYFSM 1268
sp|P43245|MDR1_RAT          AREGRTCVVIAHRLSTIQNADLIVVIQNGQVKEHGTHQQLLAQKGIYFSM 1269
tr|Q91586|Q91586_XENLA      ARMGRTCIVIAHRLSTIQNADKIAVIQNGKVVEQGTHQQLLQLKGVYFSL 1279
                            ** ****:************* *.*::**:: * ***:***  **:**::

tr|C0KKU9|C0KKU9_CANFA      VSVQAGAKR- 1281
tr|A2VBC7|A2VBC7_SHEEP      VSVQAGTKRQ 1285
sp|P08183|MDR1_HUMAN        VSVQAGTKRQ 1280
tr|Q6UUW3|Q6UUW3_RABIT      VSVQAGGKRQ 1279
sp|P21448|MDR1_CRIGR        VSVQAGAKR- 1276
sp|P06795|MDR1_MOUSE        V--QAGAKRS 1276
sp|P43245|MDR1_RAT          V--QAGAKRS 1277
tr|Q91586|Q91586_XENLA      VTIQLGHS-- 1287
                            *  * * .
```

***Figure 13** – P-gp Sequence alignment between different species: dog, frog, hamster, human, mouse, rabbit, rat and sheep. An * (asterisk) indicates positions which have a single, fully conserved residue. A : (colon) indicates conservation between groups of strongly similar properties (same color group). A . (period) indicates conservation between groups of weakly similar properties (similar shapes) [10].*

The residue colors according to their physicochemical properties:

Residue	***Color***	***Property***
AVFPMILW	*RED*	*Small (small+ hydrophobic (incl.aromatic -Y))*
DE	*BLUE*	*Acidic*
RK	*MAGENTA*	*Basic - H*
STYHCNGQ	*GREEN*	*Hydroxyl + sulfhydryl + amine + G*
Others	*Grey*	*Unusual amino/imino acids etc*

B. Sequence Alignment and Variability

For the purpose of sequence alignment and variability calculations tools from the Bioinformatics Resource Portal ExPASy were used [23]. With ClustalW version 2.1 a sequence alignment between different species considering the whole sequences as well as the transmembrane domains only was calculated. The assignment of the transmembrane regions for P-gp were taken from UniProt [24]. Further the variability of the aligned animal sequences and the mouse P-gp structure was checked with the Protein Variability Server (PVS) [25]. For all calculations the settings were used as default.

C. Homology Model

For the homology model the mouse P-gp structure (PDB ID: 3G5U, resolution: 3,8Å [7]) was taken as the template and the rat P-gp sequence was defined as the target. A sequence alignment was performed with ClustalW version 2.1. The resulting alignment was identical to the multiple sequence alignment mentioned above and used for model building with MODELLER version 9.8 [26]. All settings were kept as default. To adjust the disruption in TM helix 12 (residues 982-1000), this part was replaced by the homologous part of TM helix 6 (residues 339-357) according to Pajeva et al. [27]. To analyze the quality of the model the outliers were checked in MOE [28] and with PROCHECK [29, 30]. From the 100 generated models, the final one was chosen regarding the generously allowed and disallowed outliers, the DOPE score, Z-score, QMEAN and dfire-energy, all calculated with SWISS-MODEL [31, 32].

D. Database Search

The search for rat P-gp ligands was carried out in the Transporter Database TP search [33], in the ChEMBL database (ChEMBLdb) [34], and in PubMed [35].

E. Docking

For the docking study 6 rat P-gp inhibitors with known IC_{50} values were chosen [36]. Minimization and protonation of the ligands as well as the correct determination of ASN/GLN/HIS flips for the protein was performed with MOE. For the docking process GOLD Suite version 5.1 was utilized [19, 20]. With GOLD, hydrogens were added, the binding site was defined as the entire TM region and all side chains were kept rigid. For the calculation 100 genetic algorithm runs per molecule were performed and the scoring function GOLDScore as in GOLD implemented was used to evaluate the received complexes.

IV. Results and Discussion

A. Sequence Alignment

For the multiple sequence alignment the P-gp sequences of dog, frog, hamster, human, mouse, rabbit, rat and sheep were considered. The alignment was calculated twice with ClusalW: first the whole P-gp sequence was utilized and second only the transmembrane domains (TMDs) as described within UniProt were taken for the calculation. The results were not surprising: **Tables 3 and 4** show that the sequence similarities are very high among these species, especially between mouse and rat (92% or rather 88%). The little differences in percentages involving the whole sequence and the transmembrane domains only are expected, as the nucleotide binding domain is strongly conserved and thus the whole protein comparison shows slightly higher values than the TMD only.

Table 3 *– Results of the multiple sequence alignment of the whole P-gp sequences, shown in percent*

species	*dog*	*frog*	*hamster*	*human*	*mouse*	*rabbit*	*rat*	*sheep*
dog	*100*	*66*	*87*	*90*	*80*	*85*	*79*	*87*
frog		*100*	*68*	*67*	*63*	*66*	*63*	*65*
hamster			*100*	*87*	*82*	*85*	*82*	*84*
human				*100*	*80*	*86*	*79*	*87*
mouse					*100*	*78*	*92*	*78*
rabbit						*100*	*77*	*83*
rat							*100*	*77*
sheep								*100*

Table 4 – *Results of the multiple sequence alignment of the P-gp transmembrane domains, shown in percent*

species	dog	frog	hamster	human	mouse	rabbit	rat	sheep
dog	100	60	84	88	75	82	74	87
frog		100	59	62	57	58	58	60
hamster			100	84	77	82	76	82
human				100	77	85	76	86
mouse					100	73	88	75
rabbit						100	73	83
rat							100	74
sheep								100

B. Sequence Variability

Further the variability of the aligned animal P-gp sequences and the mouse P-gp structure was checked with the Protein Variability Server (PVS). **Figure 14** shows that the variability is higher in the beginning than at the end of the P-gp sequence. In contrast in the transmembrane regions in the middle of the sequence it is low or even not existing. For a better imagination **Figure 15** illustrates the conservation color-coded: blue represents conserved regions and red variable ones. Again the very high sequence conservation is demonstrated.

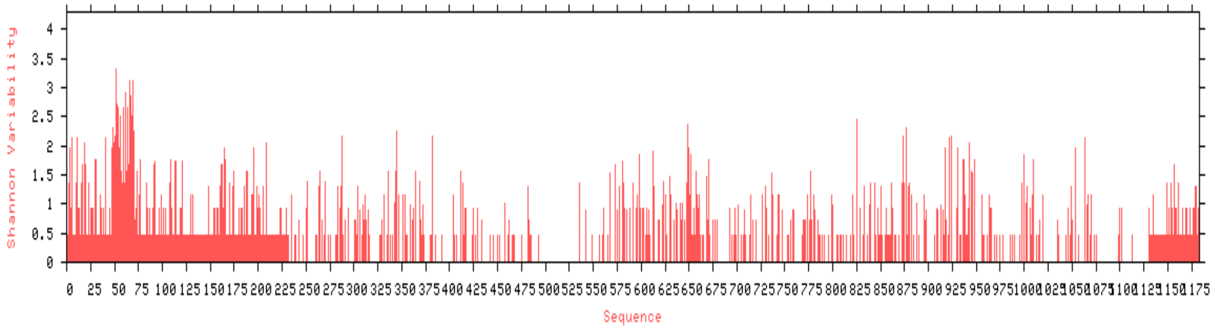


Figure 14 – *Diagram showing the variability; the higher the red peak, the higher the variability of the amino acids.*

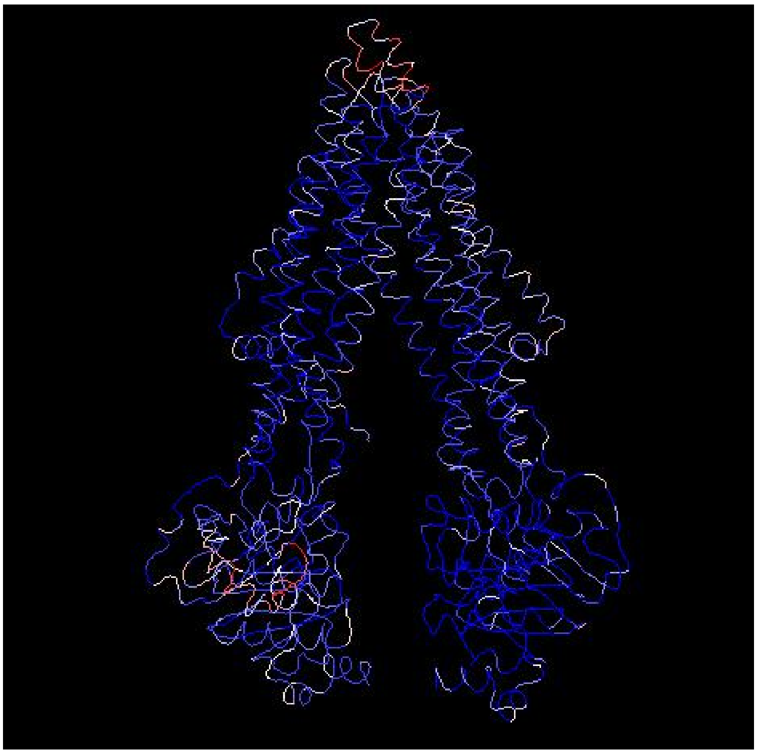


Figure 15 *– Colored P-gp illustrates the conservation; blue represents conserved regions, red variable ones.*

C. Homology Model

The publications of the mouse P-gp structure [7], the human P-gp homology model [37], and the alignment results described earlier paved the way for developing a rat P-gp model. The rat homology model was based on the structure of mouse P-gp (PDB ID: 3G5U) and its alignment to the rat P-gp sequence (sequence similarity: 92%). With these inputs the modeling program MODELLER [26] generated 100 different homology models which were subsequently refined due to the bad molpdf values (10401-11979). The deletion of a loop did only slightly improve the score. The low score was mainly due to a disruption in TM helix 12 [**Figure 16**]. This could be remarkably improved when following the procedure from Pajeva et al., by exchanging it with TM helix 6 (residues 339-357) [27].

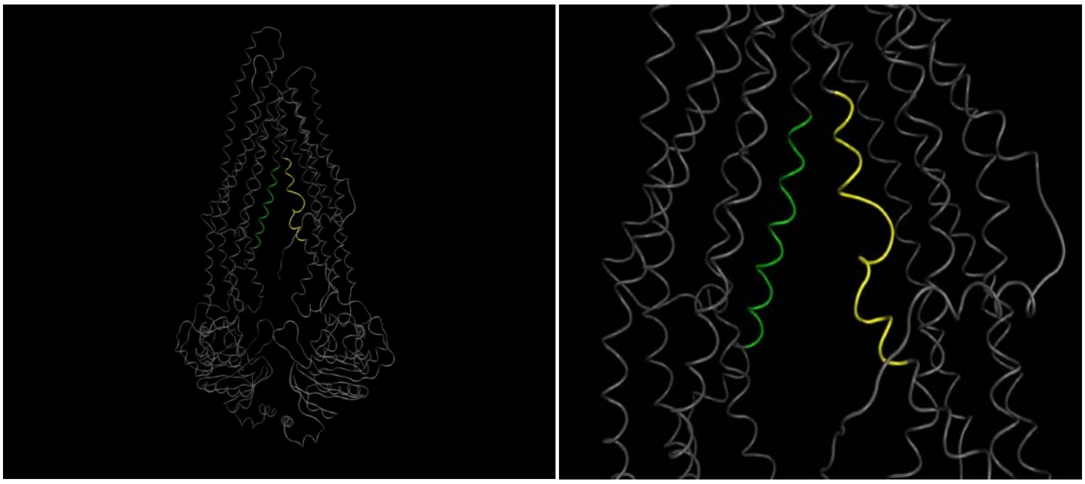
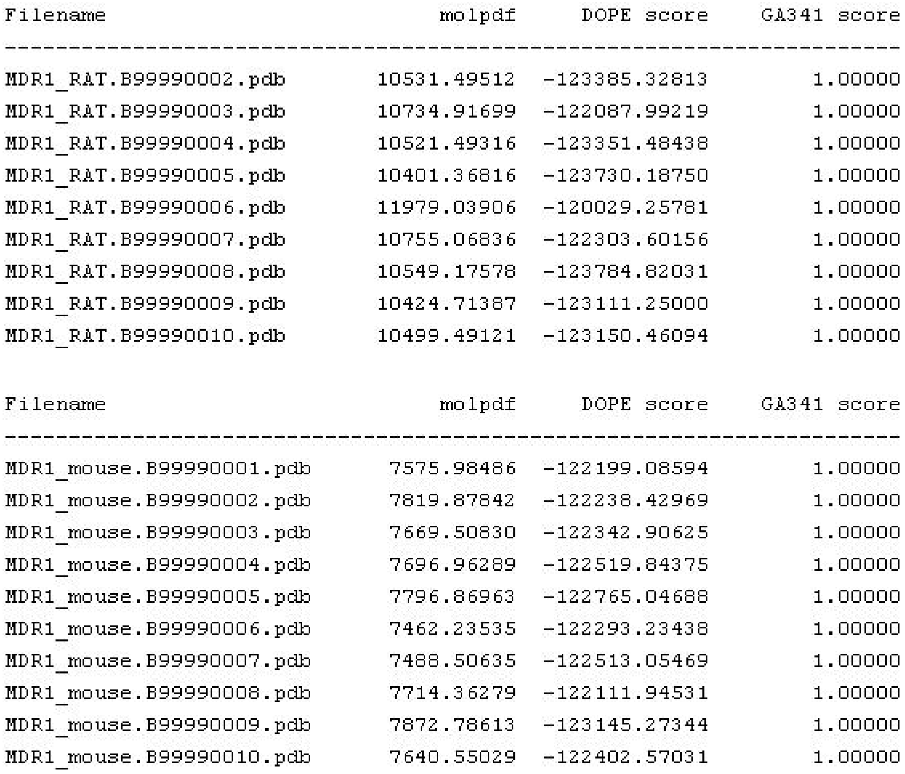


__Figure 16__ – Depiction of the disruption in TM helix 12 (yellow) and TM helix 6 (green).

The molpdf decreased substantially after this exchange, now ranging from 7462 to 7873, as shown in **Table 5**.



```
Filename                           molpdf     DOPE score    GA341 score
-----------------------------------------------------------------------
MDR1_RAT.B99990002.pdb        10531.49512  -123385.32813        1.00000
MDR1_RAT.B99990003.pdb        10734.91699  -122087.99219        1.00000
MDR1_RAT.B99990004.pdb        10521.49316  -123351.48438        1.00000
MDR1_RAT.B99990005.pdb        10401.36816  -123730.18750        1.00000
MDR1_RAT.B99990006.pdb        11979.03906  -120029.25781        1.00000
MDR1_RAT.B99990007.pdb        10755.06836  -122303.60156        1.00000
MDR1_RAT.B99990008.pdb        10549.17578  -123784.82031        1.00000
MDR1_RAT.B99990009.pdb        10424.71387  -123111.25000        1.00000
MDR1_RAT.B99990010.pdb        10499.49121  -123150.46094        1.00000

Filename                           molpdf     DOPE score    GA341 score
-----------------------------------------------------------------------
MDR1_mouse.B99990001.pdb       7575.98486  -122199.08594        1.00000
MDR1_mouse.B99990002.pdb       7819.87842  -122238.42969        1.00000
MDR1_mouse.B99990003.pdb       7669.50830  -122342.90625        1.00000
MDR1_mouse.B99990004.pdb       7696.96289  -122519.84375        1.00000
MDR1_mouse.B99990005.pdb       7796.86963  -122765.04688        1.00000
MDR1_mouse.B99990006.pdb       7462.23535  -122293.23438        1.00000
MDR1_mouse.B99990007.pdb       7488.50635  -122513.05469        1.00000
MDR1_mouse.B99990008.pdb       7714.36279  -122111.94531        1.00000
MDR1_mouse.B99990009.pdb       7872.78613  -123145.27344        1.00000
MDR1_mouse.B99990010.pdb       7640.55029  -122402.57031        1.00000
```

__Table 5__ – MODELLER scores before (above) and after (below) the exchange of the TM helix 12. The molpdf decreased substantially after exchanging the helices.

The resulting models were evaluated with the geometry check tool implemented in MOE [28]. All models were assessed with the highest possible GA341 score of 1. Additionally, the models were analyzed with PROCHECK [29, 30] and according to the obtained results the six best models (model number 77, 95, 60, 52, 21 and 54) were chosen for further validation with SWISS-MODEL [31, 32]. These tests showed in average a DOPE of -123286, a Z-Score of -401, a QMEAN of 0.4, a dfire-energy of -1513.34 and a disallowed outliers score of 0.82 [**Table 6**].

Table 6 *– Results of the model analysis with PROCKECK and SWISS-MDOEL.*

NR.	*DOPE*[a]	*disallowed*[b]	*Z-Score*[c]	*QMEAN*[d]	*dfire-energy*[e]
77	*-125664,4219*	*0,9000*	*-3,997*	*0,402*	*-1523,43*
95	*-124369,2734*	*0,8000*	*-4,018*	*0,4*	*-1502,16*
60	*-125507,2891*	*0,9000*	*-4,134*	*0,389*	*-1515,94*
52	*-125562,9141*	*0,8000*	*-4,064*	*0,395*	*-1517,53*
21	*-125046,2656*	*0,7000*	*-3,861*	*0,414*	*-1511,35*
54	*-125562,9141*	*0,8000*	*-4,003*	*0,401*	*-1509,63*

[a] *discrete optimized energy,*
[b] *disallowed outliers in the Ramachandran plot,*
[c] *measure for the absolute quality of the model,*
[d] *score of the whole model reflecting the predicted model reliability ranging from 0 to1,*
[e] *assessment of non bonded atomic interactions [31]*

As a result of this analysis, model number 21 was chosen as the best model with a DOPE of -125046, a disallowed outliers score of 0.7, a Z-Score of -3.86, a QMEAN of 0.41 and a dfire-energy of -1511 [**Figure 17**].

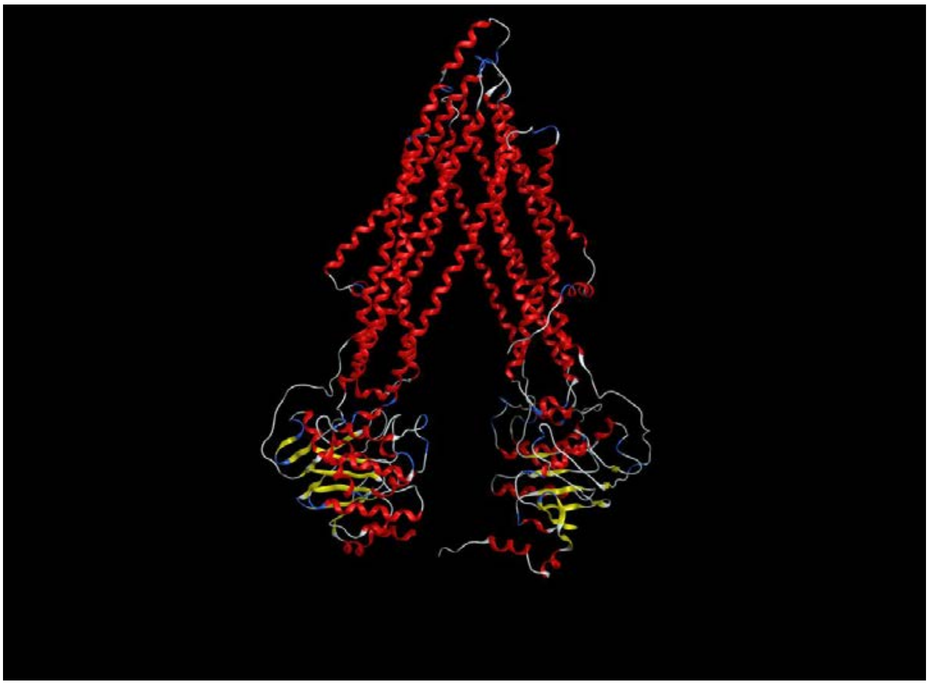


Figure 17 *– Depiction of the rat P-gp model number 21.*

Subsequently, the models were color coded according to their similarities and differences in the amino acid sequence. **Figure 18** shows that not only as already mentioned before the sequence similarity between species is high but also the sequence identity.

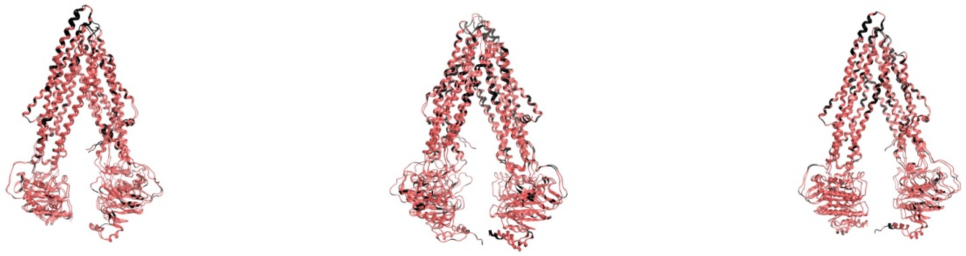


Figure 18 *– Comparison of the human, mouse and rat P-gps, pink demonstrate identical amino acids, black shows different amino acids, (a) human and mouse P-gp, (b) human and rat P-gp, (c) mouse and rat P-gp.*

D. Database search

In a next step we focused on the search for substrates and inhibitors of the human, mouse and rat P-gps. Especially rat P-gp ligands were of interest in order to carry out docking studies. For this purpose two databases were consulted: the Transport Database [33] and the ChEMBL database (ChEMBLdb) [34].

Surprisingly, this task turned out to be a tricky undertaking, as no rat P-gp ligands were found in these databases. On the other hand, numerous human and mouse P-gp substrates and inhibitors were found twice, threefold, etc. in one database, which made data collection very elaborative and time consuming. Nevertheless, in the end it was possible to filter out the requested information.

1. Transporter Database

In the Transporter Database only human substrates and inhibitors and mouse substrates were found. There was no information for the rat at all. 256 human substrates and 12 mouse substrates were detected, whereupon all 12 mouse substrates overlap with the human substrates. On the other hand, only 371 human but no mouse or rat inhibitors were retrieved [**Figure 19 and 20**].

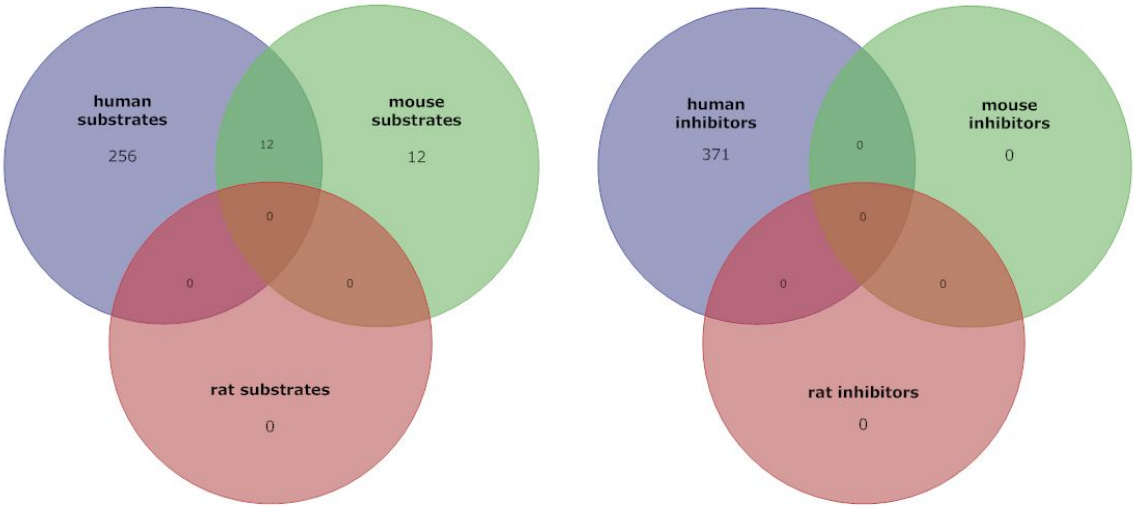


Figure 19 *– Venn diagrams of the P-gp ligands found in the TP-database [33].*

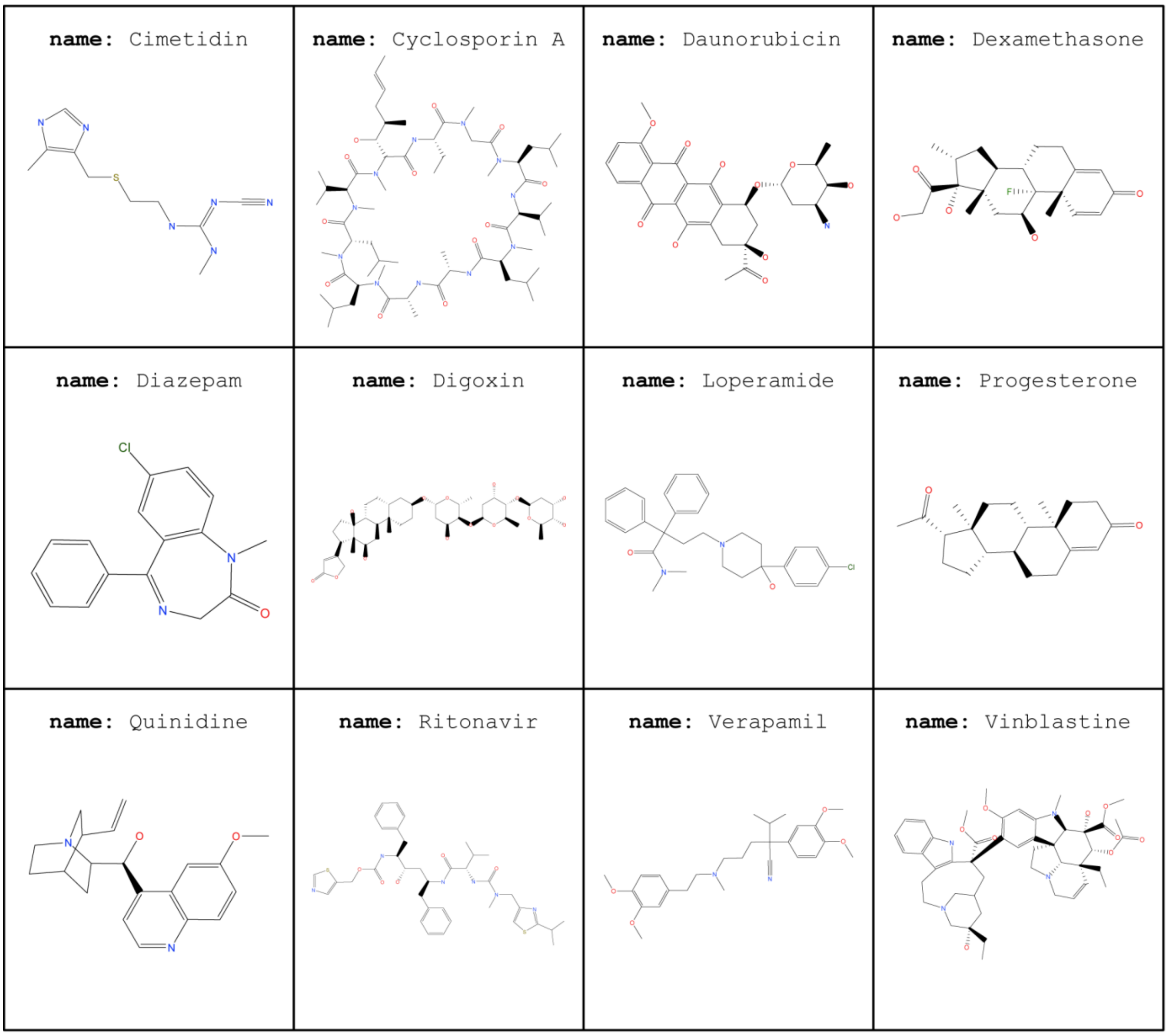

name: Cimetidin
name: Cyclosporin A
name: Daunorubicin
name: Dexamethasone
name: Diazepam
name: Digoxin
name: Loperamide
name: Progesterone
name: Quinidine
name: Ritonavir
name: Verapamil
name: Vinblastine

Figure 20 *– Overlapping 12 human and mouse substrates retrieved by Tp-search [33].*

2. ChEMBL Database

The ChEMBLdb [34] is not sub classified into substrates and inhibitors like the TP search database. One can only find ligands which are described more precisely in their profile. During the request once again only human and mouse ligands, but not a single rat ligand was returned [**Figure 21 and 22**]. In total 1087 human and 110 mouse ligands were detected, with 33 overlapping ones.

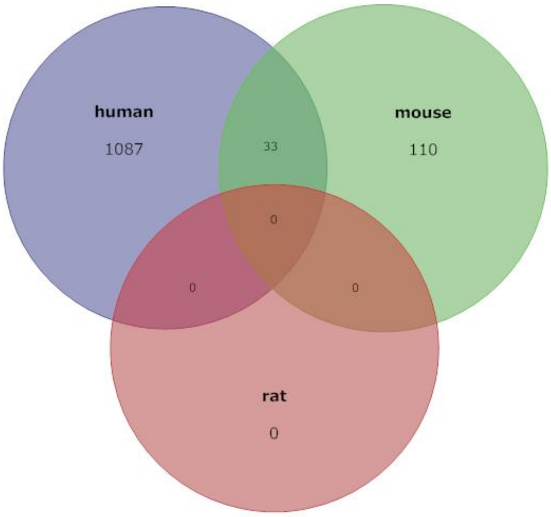
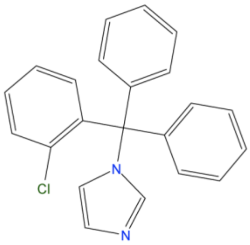
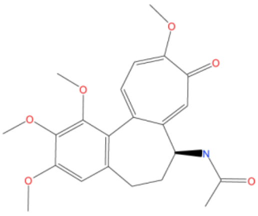
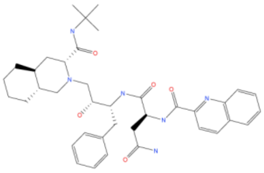
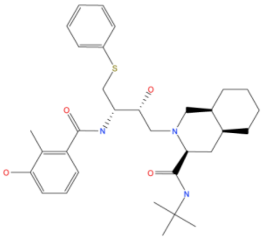
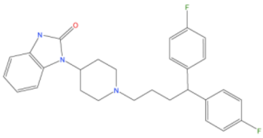
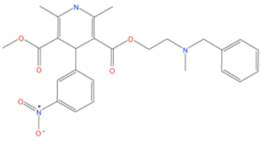
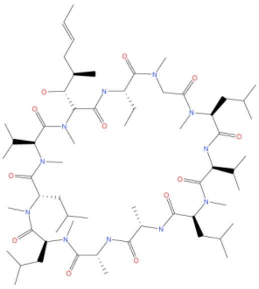
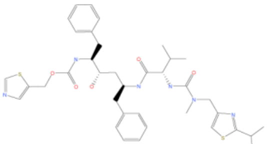
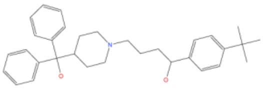
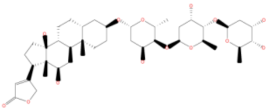
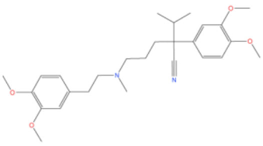
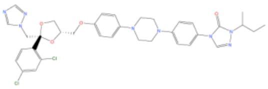
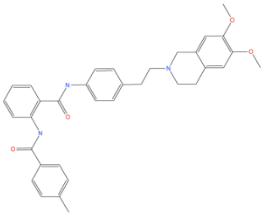
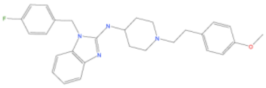
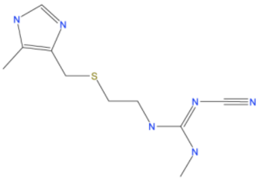
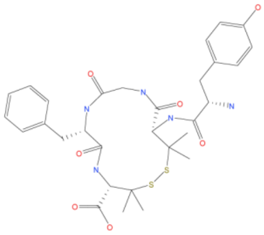
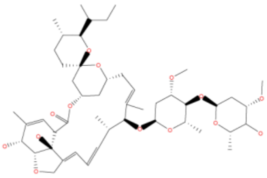
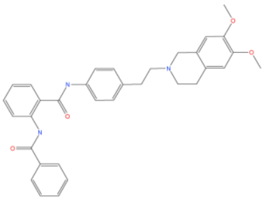
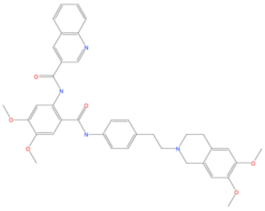
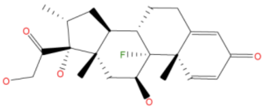
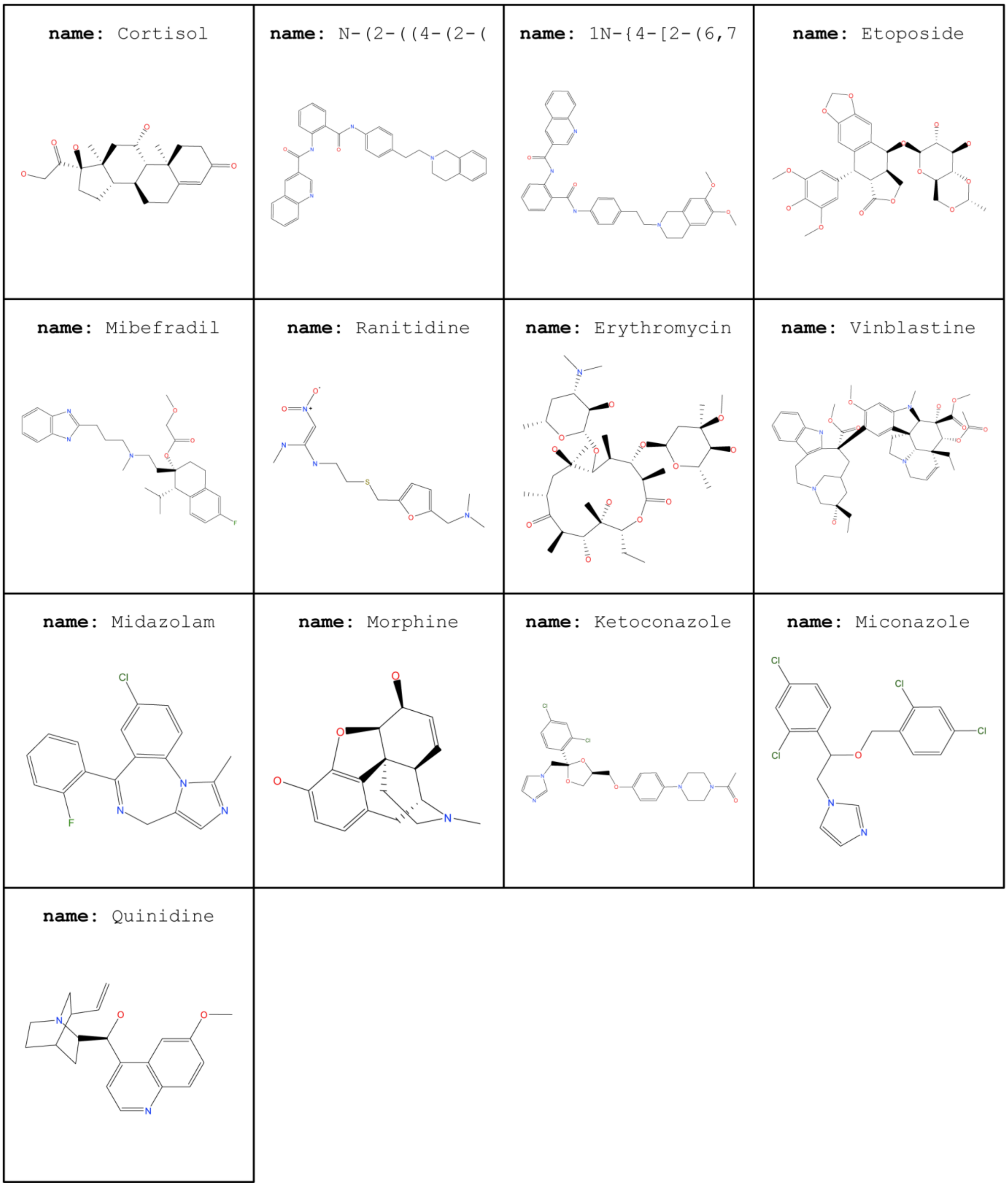


Figure 21 *– Venn diagram of the P-gp ligands detected in the ChEMBLdb [34] for different species.*

name: Clotrimazole 	**name:** Colchicine 	**name:** Saquinavir 	**name:** Nelfinavir 
name: Pimozide 	**name:** Nicardipine 	**name:** Cyclosporin A 	**name:** Ritonavir 
name: (+/-)1-(4-ter 	**name:** Digoxin 	**name:** Verapamil 	**name:** Itraconazole 
name: XR9504 	**name:** Astemizole 	**name:** Cimetidine 	**name:** (4S,7S,13S)-1 
name: Ivermectin 	**name:** 1N-{4-[2-(6,7 	**name:** Tariquidar 	**name:** Dexamethasone 



name: Cortisol

name: N-(2-((4-(2-(

name: 1N-{4-[2-(6,7

name: Etoposide

name: Mibefradil

name: Ranitidine

name: Erythromycin

name: Vinblastine

name: Midazolam

name: Morphine

name: Ketoconazole

name: Miconazole

name: Quinidine

Figure 22 *– Overlapping 33 human and mouse compounds of the ChEMBL search[34].*

3. PubMed

Having not found any rat data in the main public available databases, in the next step a literature search in PubMed [35] was conducted. A few articles were detected leading to in total 18 substrates and inhibitors of rat P-gp [**Figure 23**] [36, 38-43].

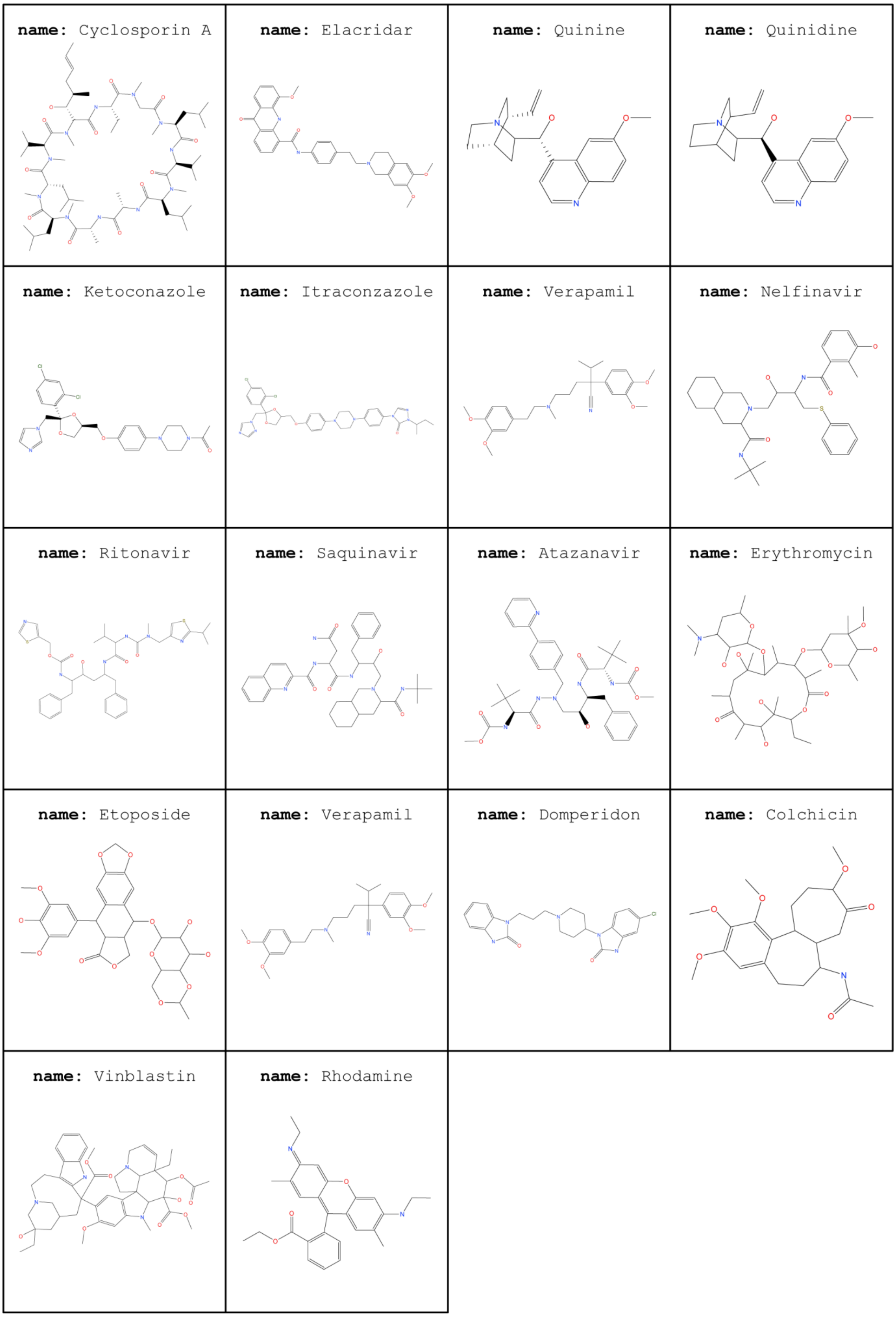

name: Cyclosporin A
name: Elacridar
name: Quinine
name: Quinidine
name: Ketoconazole
name: Itraconzazole
name: Verapamil
name: Nelfinavir
name: Ritonavir
name: Saquinavir
name: Atazanavir
name: Erythromycin
name: Etoposide
name: Verapamil
name: Domperidon
name: Colchicin
name: Vinblastin
name: Rhodamine

***Figure 23** – Resulting 18 rat P-gp ligands returned from PubMed [35].*

E. Docking

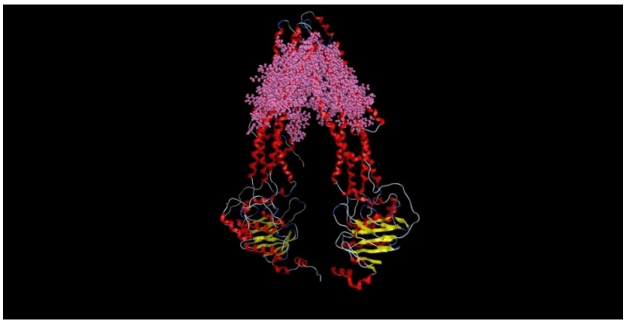


Figure 24 *– Binding site of the P-gp marked in purple.*

The docking software GOLD [19, 20] was used to dock 6 rat P-gp inhibitors [36] (taken from PubMed search) with known IC_{50} values against the rat homology model. The minimization and protonation of the ligands as well as the correct determination of ASN/GLN/HIS flips for the protein was performed with MOE [28]. According to Klepsch et al. [37] there is evidence that the proteins pore is filled with water and therefore it was suggested to dock the ligands in their ionized state. Before the docking run hydrogens were added to the protein using GOLD and the binding site was defined as the entire TM region [**Figure 24**]. During docking all side chains were kept rigid. GOLD is based on a genetic algorithm and for each of the 6 ligands 100 docking poses were calculated. Subsequently, the scoring function GOLDScore implemented in GOLD was used to rank the complexes.

The obtained poses were located in the whole TM region,showing interactions with residues of different TM helices. The analysis of the complexes showed that especially TM helices 1, 5, 6 and 12 were involved in interactions. Frequently, interactions were observed with residues F70 located in TM helix 1 and F335 in TM helix 6. Also residue T306 was involved in some interactions, whose human analogue T307 was shown to be important in ligand interactions [**Figure 25**] [37].

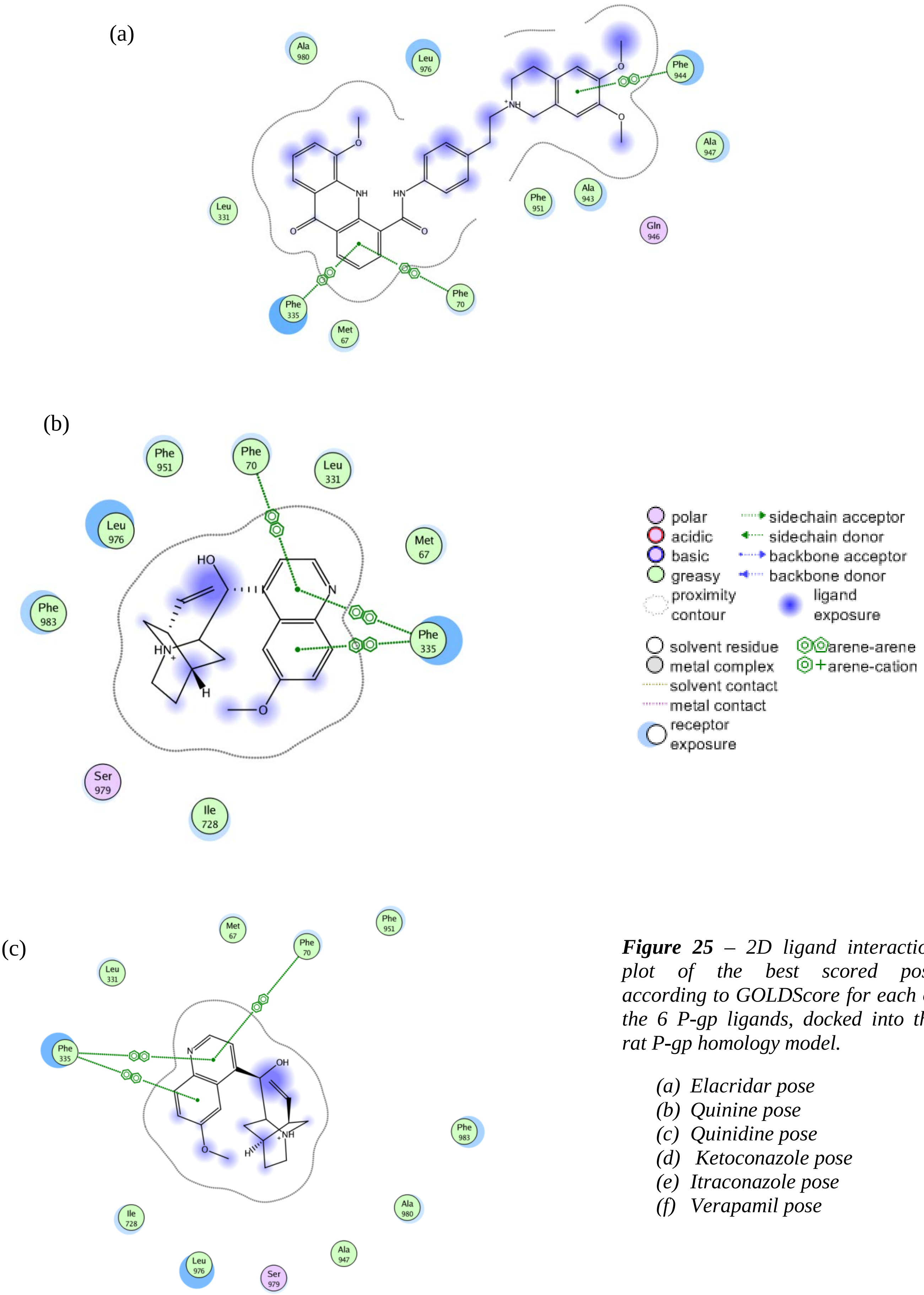
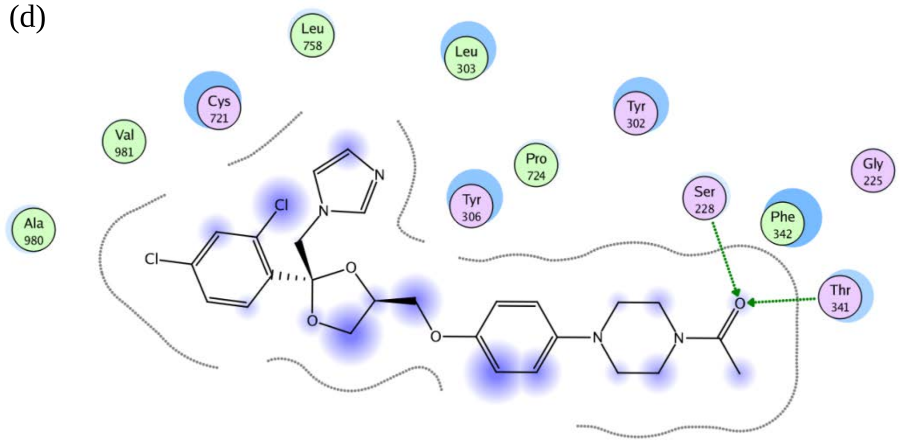

(a)
Ala 980
Leu 976
Phe 944
Ala 947
Leu 331
Phe 951
Ala 943
Gln 946
Phe 335
Met 67
Phe 70
NH
HN
+NH
(b)
Phe 951
Phe 70
Leu 331
Leu 976
Met 67
Phe 983
Phe 335
Ser 979
Ile 728
HO
HN+
H
N
O
polar
acidic
basic
greasy
proximity contour
sidechain acceptor
sidechain donor
backbone acceptor
backbone donor
ligand exposure
solvent residue
metal complex
solvent contact
metal contact
receptor exposure
arene-arene
arene-cation
(c)
Met 67
Phe 951
Phe 70
Leu 331
Phe 335
Phe 983
Ile 728
Ala 980
Ala 947
Leu 976
Ser 979
N
OH
+NH
H

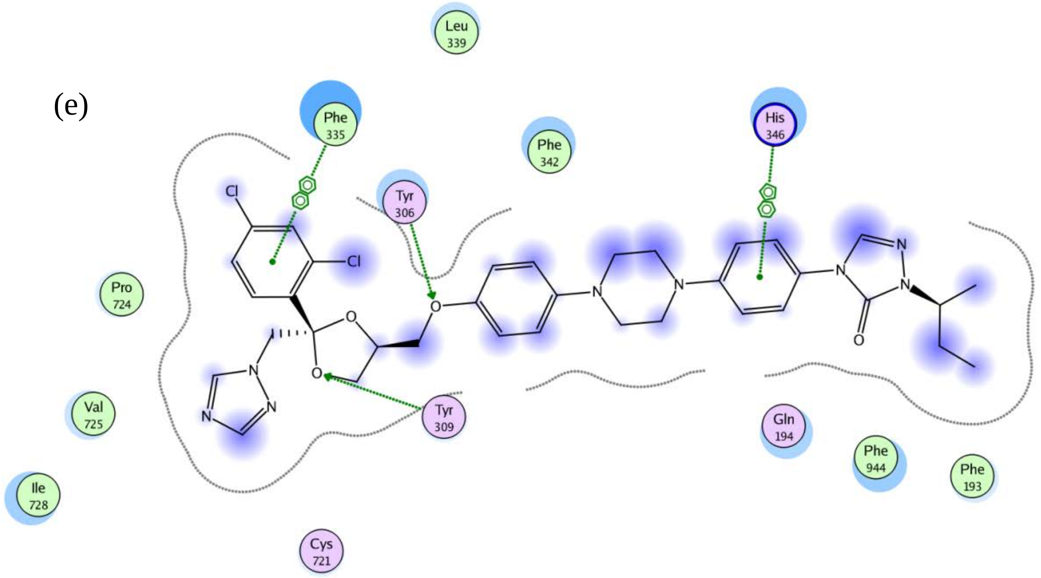
Figure 25 *– 2D ligand interaction plot of the best scored pose according to GOLDScore for each of the 6 P-gp ligands, docked into the rat P-gp homology model.*

- *(a) Elacridar pose*
- *(b) Quinine pose*
- *(c) Quinidine pose*
- *(d) Ketoconazole pose*
- *(e) Itraconazole pose*
- *(f) Verapamil pose*

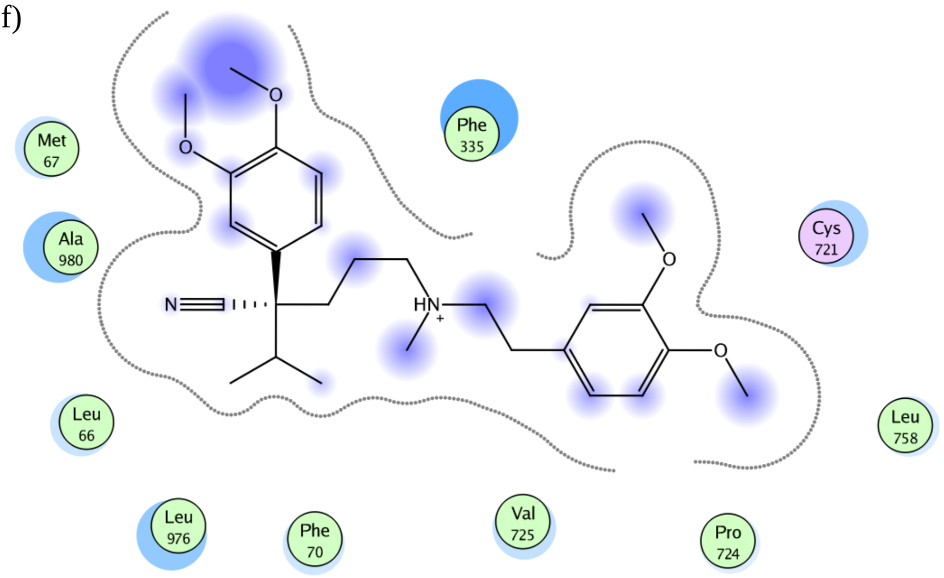
(d)


Leu 758
Leu 303
Cys 721
Val 981
Tyr 302
Pro 724
Gly 225
Ala 980
Tyr 306
Ser 228
Phe 342
Thr 341
Cl
Cl
N
N
O
O
O
N
N
O

(e)


Leu 339
Phe 335
Phe 342
His 346
Tyr 306
Cl
Cl
Pro 724
O
O
O
N
N
N
N
N
N
N
N
O
Tyr 309
Val 725
Ile 728
Gln 194
Phe 944
Phe 193
Cys 721

(f)


O
O
Phe 335
Met 67
Ala 980
Cys 721
N
HN+
O
O
Leu 66
Leu 758
Leu 976
Phe 70
Val 725
Pro 724

A representative docking pose is depicted in **Figure 26** showing the best ranked Chinine receptor complex.

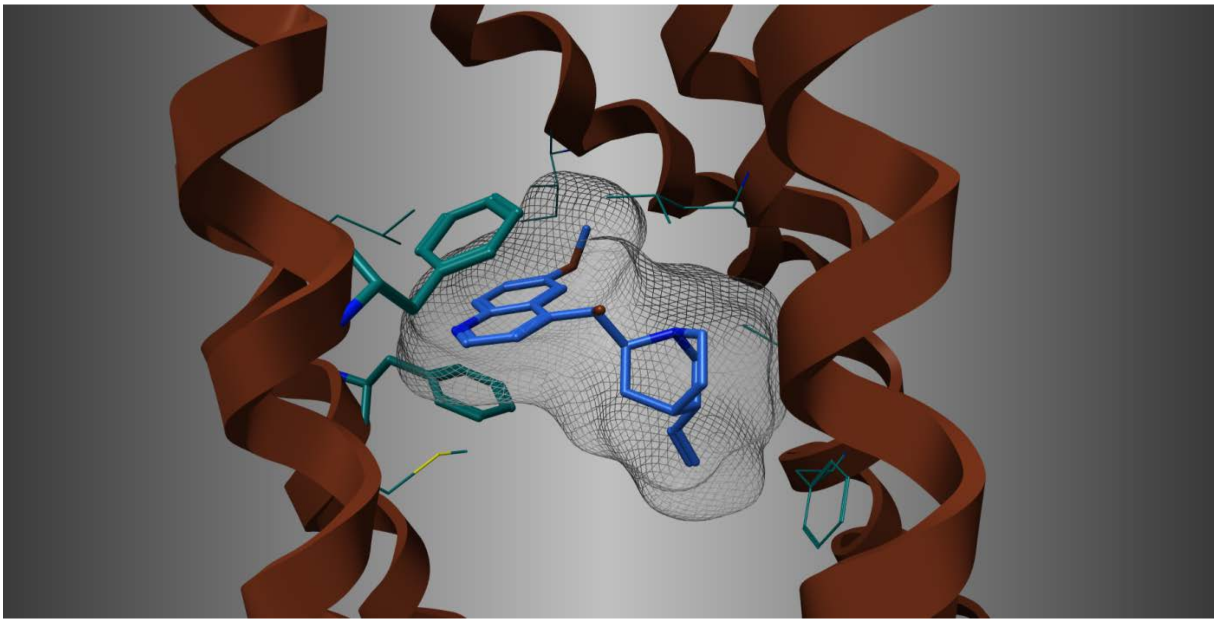


***Figure 26** – Best scored Chinine pose according to GOLDScore.*

Finally, the docking results were compared with the experimental IC_{50} values from Zolnerciks et al. [36]. The docking ranks were obtained according to the Fitness scores from the "bestranking" file produced by GOLD. Surprisingly, the experimental and docking ranks were almost identical disregarding only ranks 3 and 4 which were switched [**Table 7**].

***Table 7** – Comparison of rankings according to experimental data and GOLD scoring*

Inhibitor	*Rank (exper)*	*Rank (dock)*
Elacridar	*1*	*1*
Quinine	*6*	*6*
Quinidine	*5*	*5*
Ketoconazole	*4*	*3*
Itraconazole	*2*	*2*
Verapamil	*3*	*4*

V. Summary and Outlook

In general the drug development workflow can be divided into 3 main stages: drug discovery, preclinical stage (animal in vivo trials) and clinical stage (human in vivo trials). Especially in the early drug discovery stage in silico predictions are widely applied. One important field of *in silico* activity is the generation of reliable models for ADME and toxicity profiles. Here P-gp plays a key role because of its biological function as a xenobiotic carrier between distribution compartments.

Nowadays a huge amount of experimental data against human P-gp is already available and has been implemented in the generation of prediction models during the drug discovery stage. However, in the early preclinical phase of animal in vivo studies, the animal P-gp activity profile may differ significantly and may lead to attrition. Thus next to existing *in silico* predictive models against human P-gp activity, predictive *in silico* models against rat and mouse would be beneficial.

In our study we tackled this by the generation of a structure based rat P-gp prediction tool. Due to the lack of a crystal structure of rat P-gp, homology modeling and computational ligand docking represent the only possibilities for structure-based hypotheses for protein-ligand-interactions. Therefore, the accurate prediction of membrane protein structures and their interaction with small molecules stays a challenge.

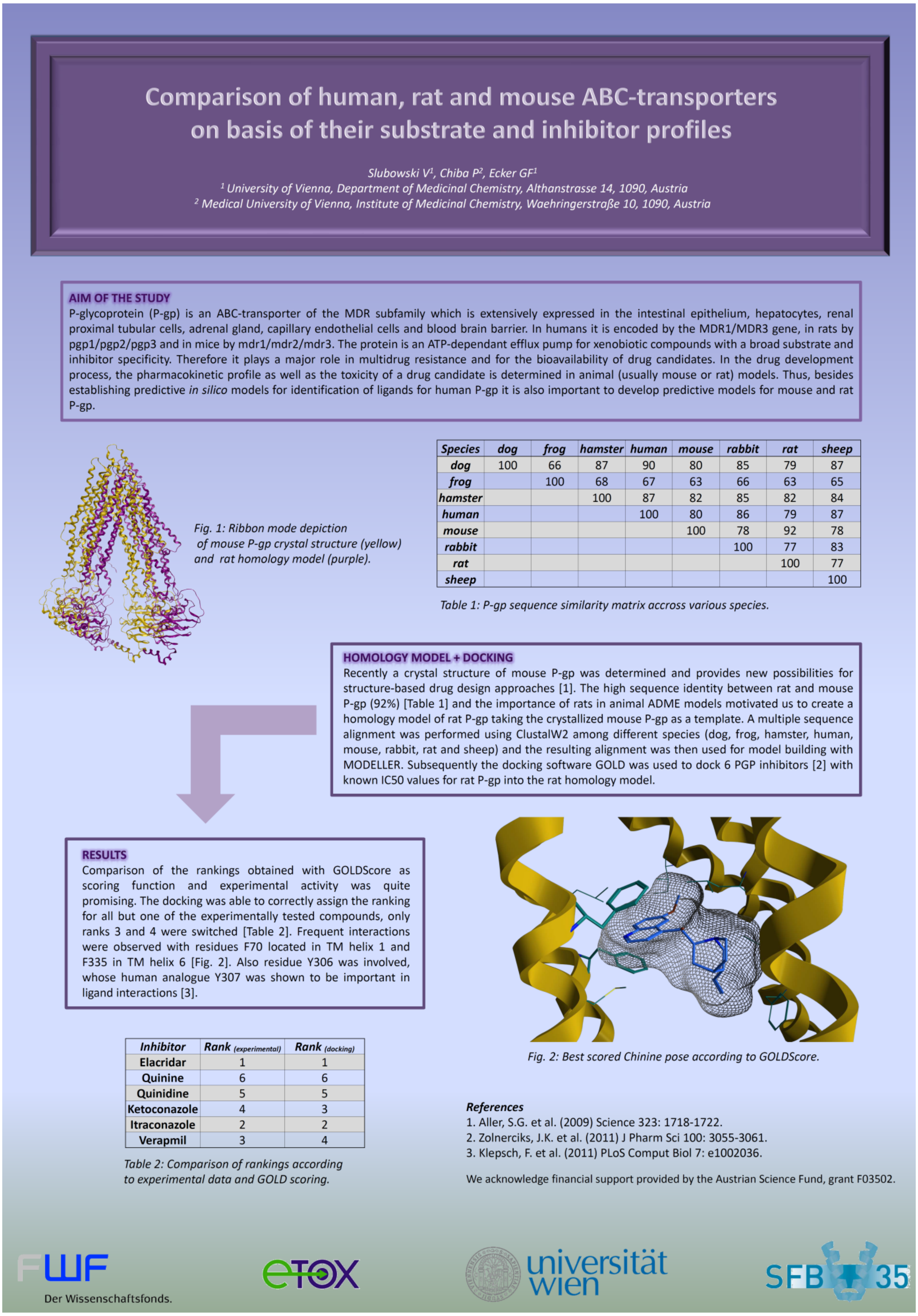
With our work we tried to take the first step towards *in silico* ADME and toxicity predictions with a focus on the role of P-gp. This would allow to assess potential failure of a drug candidate at an early stage in the drug development pipeline. For this reason we first constructed a homology model of rat P-gp. Then we compiled a ligand library composed of known rat P-gp ligands from literature. A subset of this library was then docked into the homology model and the subsequent ranking list was compared to the experimental (IC_{50}) rankings.

The resulted ranking list was promising: the docking was able to correctly assign almost all ranks, only ranks 3 and 4 were switched. Of course additional validation needs to be done, but the obtained results in this study assume its suitability for structure-based prediction models.

However, we have to consider that the amount of the docked compounds for validation was limited. For a more sophisticated validation more experimental data on compounds from industry is needed.

Next to the predictive capabilities, the structural insights of the complexes can as well ameliorate our undertanding and hypothesis of inhibitor binding on a molecular level, stimulating scientists to conduct new experiments.

Poster



Comparison of human, rat and mouse ABC-transporters on basis of their substrate and inhibitor profiles

Slubowski V[1], Chiba P[2], Ecker GF[1]
[1] University of Vienna, Department of Medicinal Chemistry, Althanstrasse 14, 1090, Austria
[2] Medical University of Vienna, Institute of Medicinal Chemistry, Waehringerstraße 10, 1090, Austria

AIM OF THE STUDY

P-glycoprotein (P-gp) is an ABC-transporter of the MDR subfamily which is extensively expressed in the intestinal epithelium, hepatocytes, renal proximal tubular cells, adrenal gland, capillary endothelial cells and blood brain barrier. In humans it is encoded by the MDR1/MDR3 gene, in rats by pgp1/pgp2/pgp3 and in mice by mdr1/mdr2/mdr3. The protein is an ATP-dependant efflux pump for xenobiotic compounds with a broad substrate and inhibitor specificity. Therefore it plays a major role in multidrug resistance and for the bioavailability of drug candidates. In the drug development process, the pharmacokinetic profile as well as the toxicity of a drug candidate is determined in animal (usually mouse or rat) models. Thus, besides establishing predictive *in silico* models for identification of ligands for human P-gp it is also important to develop predictive models for mouse and rat P-gp.

Fig. 1: Ribbon mode depiction of mouse P-gp crystal structure (yellow) and rat homology model (purple).

Species	*dog*	*frog*	*hamster*	*human*	*mouse*	*rabbit*	*rat*	*sheep*
dog	100	66	87	90	80	85	79	87
frog		100	68	67	63	66	63	65
hamster			100	87	82	85	82	84
human				100	80	86	79	87
mouse					100	78	92	78
rabbit						100	77	83
rat							100	77
sheep								100

Table 1: P-gp sequence similarity matrix accross various species.

HOMOLOGY MODEL + DOCKING

Recently a crystal structure of mouse P-gp was determined and provides new possibilities for structure-based drug design approaches [1]. The high sequence identity between rat and mouse P-gp (92%) [Table 1] and the importance of rats in animal ADME models motivated us to create a homology model of rat P-gp taking the crystallized mouse P-gp as a template. A multiple sequence alignment was performed using ClustalW2 among different species (dog, frog, hamster, human, mouse, rabbit, rat and sheep) and the resulting alignment was then used for model building with MODELLER. Subsequently the docking software GOLD was used to dock 6 PGP inhibitors [2] with known IC50 values for rat P-gp into the rat homology model.

RESULTS

Comparison of the rankings obtained with GOLDScore as scoring function and experimental activity was quite promising. The docking was able to correctly assign the ranking for all but one of the experimentally tested compounds, only ranks 3 and 4 were switched [Table 2]. Frequent interactions were observed with residues F70 located in TM helix 1 and F335 in TM helix 6 [Fig. 2]. Also residue Y306 was involved, whose human analogue Y307 was shown to be important in ligand interactions [3].

Fig. 2: Best scored Chinine pose according to GOLDScore.

Inhibitor	*Rank (experimental)*	*Rank (docking)*
Elacridar	1	1
Quinine	6	6
Quinidine	5	5
Ketoconazole	4	3
Itraconazole	2	2
Verapmil	3	4

Table 2: Comparison of rankings according to experimental data and GOLD scoring.

References
1. Aller, S.G. et al. (2009) Science 323: 1718-1722.
2. Zolnerciks, J.K. et al. (2011) J Pharm Sci 100: 3055-3061.
3. Klepsch, F. et al. (2011) PLoS Comput Biol 7: e1002036.

We acknowledge financial support provided by the Austrian Science Fund, grant F03502.

Der Wissenschaftsfonds.

List of abbreviations

ABC	ATP Binding Cassette
ADME	Absorption, Distribution, Metabolism and Excretion
ATP	Adenosine Triphosphate
ChEMBLdb	ChEMBL Database
DNA	Desoxyribonucleotide Acid
DOPE	Discrete Optimized Energy
EBI	European Bioinformatics Institute
FDA	Food and Drug Administration in the USA
GA	Genetic Algorithm
GMP	Good Manufacturing Practice
IC50	Half Maximal Inhibitory Concentration
IND	Investigational New Drug
MDR	Multidrug Resistance
Molpdf	Molecular PDF
NBD	Nucleotide Binding Domain
NDA	New Drug Application
PDB	Protein Database
P-gp	P-Glycoprotein
PVS	Protein Variability Server
RNA	Ribonucletide Acid
TAP	Antigen Peptide Transporter
TM	Transmembrane

References

1. www.fda.gov.
2. DiMasi, J.A., R.W. Hansen, and H.G. Grabowski, *The price of innovation: new estimates of drug development costs.* J Health Econ, 2003. **22**(2): p. 151-85.
3. www.phrma.org.
4. www.csdd.tufts.edu.
5. Davidson, A.L., et al., *Structure, function, and evolution of bacterial ATP-binding cassette systems.* Microbiol Mol Biol Rev, 2008. **72**(2): p. 317-64, table of contents.
6. Vasiliou, V., K. Vasiliou, and D.W. Nebert, *Human ATP-binding cassette (ABC) transporter family.* Hum Genomics, 2009. **3**(3): p. 281-90.
7. Aller, S.G., et al., *Structure of P-glycoprotein reveals a molecular basis for poly-specific drug binding.* Science, 2009. **323**(5922): p. 1718-22.
8. Kwon, Y., A.V. Kamath, and M.E. Morris, *Inhibitors of P-glycoprotein-mediated daunomycin transport in rat liver canalicular membrane vesicles.* J Pharm Sci, 1996. **85**(9): p. 935-9.
9. www.drugdevelopment-technology.com.
10. www.ebi.ac.uk.
11. Larkin, M.A., et al., *Clustal W and Clustal X version 2.0.* Bioinformatics, 2007. **23**(21): p. 2947-8.
12. www.unil.ch.
13. www.biskit.pasteur.fr.
14. Baker, D. and A. Sali, *Protein structure prediction and structural genomics.* Science, 2001. **294**(5540): p. 93-6.
15. www.salilab.org.
16. Lengauer, T. and M. Rarey, *Computational methods for biomolecular docking.* Curr Opin Struct Biol, 1996. **6**(3): p. 402-6.
17. Kitchen, D.B., et al., *Docking and scoring in virtual screening for drug discovery: methods and applications.* Nat Rev Drug Discov, 2004. **3**(11): p. 935-49.

18. Jones, G., et al., *Development and validation of a genetic algorithm for flexible docking.* J Mol Biol, 1997. **267**(3): p. 727-48.
19. www.ccdc.cam.ac.uk.
20. Verdonk, M.L., et al., *Improved protein-ligand docking using GOLD.* Proteins, 2003. **52**(4): p. 609-23.
21. Higgins, D.G., J.D. Thompson, and T.J. Gibson, *Using CLUSTAL for multiple sequence alignments.* Methods Enzymol, 1996. **266**: p. 383-402.
22. www.ebi.ac.uk/Tools/msa/clustalw2/.
23. www.expasy.org.
24. www.uniprot.org.
25. www.imed.med.ucm.es/PVS/.
26. Sali, A., *Comparative protein modeling by satisfaction of spatial restraints.* Mol Med Today, 1995. **1**(6): p. 270-7.
27. Pajeva, I.K., C. Globisch, and M. Wiese, *Comparison of the inward- and outward-open homology models and ligand binding of human P-glycoprotein.* FEBS J, 2009. **276**(23): p. 7016-26.
28. www.chemcomp.com.
29. www.ebi.ac.uk/thornton-srv/software/PROCHECK/.
30. Laskowski, R.A., V.V. Chistyakov, and J.M. Thornton, *PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids.* Nucleic Acids Res, 2005. **33**(Database issue): p. D266-8.
31. www.swissmodel.expasy.org.
32. Arnold, K., et al., *The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling.* Bioinformatics, 2006. **22**(2): p. 195-201.
33. www.tp-search.jp.
34. Gaulton, A., et al., *ChEMBL: a large-scale bioactivity database for drug discovery.* Nucleic Acids Res, 2012. **40**(1): p. D1100-7.
35. www.ncbi.nlm.nih.gov/pubmed/.

36. Zolnerciks, J.K., et al., *Substrate- and species-dependent inhibition of P-glycoprotein-mediated transport: implications for predicting in vivo drug interactions.* J Pharm Sci, 2011. **100**(8): p. 3055-61.

37. Klepsch, F., P. Chiba, and G.F. Ecker, *Exhaustive sampling of docking poses reveals binding hypotheses for propafenone type inhibitors of P-glycoprotein.* PLoS Comput Biol, 2011. **7**(5): p. e1002036.

38. Takeuchi, T., et al., *Marked impact of P-glycoprotein on the absorption of TAK-427 in rats.* Biopharm Drug Dispos, 2008. **29**(6): p. 311-23.

39. Piao, Y.J., X. Li, and J.S. Choi, *Effects of verapamil on etoposide pharmacokinetics after intravenous and oral administration in rats.* Eur J Drug Metab Pharmacokinet, 2008. **33**(3): p. 159-64.

40. Syvanen, S., et al., *Pharmacokinetics of P-glycoprotein inhibition in the rat blood-brain barrier.* J Pharm Sci, 2008. **97**(12): p. 5386-400.

41. Dan, Y., et al., *Distribution of domperidone into the rat brain is increased by brain ischaemia or treatment with the P-glycoprotein inhibitor verapamil.* J Pharm Pharmacol, 2002. **54**(5): p. 729-33.

42. Cisternino, S., et al., *In vivo saturation of the transport of vinblastine and colchicine by P-glycoprotein at the rat blood-brain barrier.* Pharm Res, 2003. **20**(10): p. 1607-11.

43. Hirsch-Ernst, K.I., et al., *Inhibitors of mdr1-dependent transport activity delay accumulation of the mdr1 substrate rhodamine 123 in primary rat hepatocyte cultures.* Toxicology, 2001. **167**(1): p. 47-57.

Curriculum Vitae

Persönliche Daten

Vorname: Victoria

Zuname: Slubowski

Geburtsdatum: 26.06.1985

Familienstand: ledig

Wohnadresse: Vernholzgasse 4/9/28, 1220 Wien

Staatsbürgerschaft: Österreich

Schulische Ausbildung

1991-1995	Volkschule Institut Sacré-Coeur der Erzdiözese Wien
1995-2003	Gymnasium w.o.
2003-2005	Medizinstudium - Medizinische Universität Wien
seit 2005	Pharmaziestudium - Universität Wien
Februar 2008	1. Diplomprüfung
seit März 2010	Diplomarbeit am Department für Pharm. Chemie

Studienbegleitende Tätigkeiten

Juli 2005	Sanitätsgehilfe im Krankenhaus Lainz
Dezember 2005-2007	Reception Support bei Firmenich GesmbH
Jänner 2008-2009	Travel Coach bei Austrian Airlines
Juli/August 2009	Ferialpraktikantin bei Bayer Schering Pharma
August-September 2011	freie Dienstnehmerin bei Universität Wien
Seit März 2009	C21 Apotheke

Sonstiges

2000-2003	Vierfache Junioren-Staatsmeisterin im Rudern