Development and Application of Computational Methods in Antitubercular Drug Design

Identification of Novel Inhibitors of Ribonucleotide Reductase

DANIEL MUTHAS
Dissertation presented at Uppsala University to be publicly examined in B22, BMC, Uppsala, Friday, April 24, 2009 at 09:15 for the degree of Doctor of Philosophy (Faculty of Pharmacy). The examination will be conducted in English.

Abstract

Tuberculosis kills approximately 1.7 million people each year around the world making it one of the most lethal infectious diseases. This thesis concerns the development of two computational tools that can support the early stages of drug discovery, and their use in an anti-tubercular drug discovery program.

One of the tools developed is a statistical molecular design (SMD) approach that generates information-rich libraries biased towards a lead structure. The other method is a post-filtering technique to increase the success of virtual screening, has also been developed. Both methods have been validated using literature data.

Ribonucleotide reductase (RNR) has been identified as a potential anti-tubercular target, and our focus has been to develop small-molecule inhibitors of this target. The enzyme consists of two subunits (a large R1 and a small R2 subunit) that have to associate in order to generate a bioactive complex. It had previously been shown that a heptapeptide corresponding to the small R2 subunits C-terminal inhibited the enzyme. In order to investigate the requirements for inhibitory effect of the peptide a library was designed using the developed SMD approach. The designed library was synthesized and evaluated for biological activity and an OPLS-DA model was derived to understand which positions were most important for activity.

In order to identify small-molecule inhibitors of RNR a combined shape- and structure-based virtual screen was performed, employing ROCS, GlideXP and the developed post-filtering technique. Starting from a library of 1.5 million compounds 24 was acquired and evaluated for enzymatic activity. The best compounds were almost as potent as the starting peptide, but considerably more drug-like.

Keywords: tuberculosis, drug discovery, ribonucleotide reductase, statistical molecular design, virtual screening, pharmacophore, focused hierarchical design

Daniel Muthas, Organic Pharmaceutical Chemistry, Box 574, Uppsala University, SE-75123 Uppsala, Sweden

© Daniel Muthas 2009

ISSN 1651-6192
ISBN 978-91-554-7460-7
urn:nbn:se:uu:diva-99173 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-99173)
“Medical scientists now have both the knowledge they need to wipe out tuberculosis as a public-health problem and the tools to finish the job.”

TIME magazine, August 3rd, 1953
List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals:

I. Muthas D, Lek PM, Nurbo J, Karlén A, Lundstedt T
   *Focused hierarchical design of peptide libraries – follow the lead.* J. Chemom., 2007, 21, 486-495

II. Nurbo J, Roos AK, Muthas D, Wahlström E, Ericsson DJ, Lundstedt T, Unge T, Karlén A

III. Muthas D, Sabnis YA, Lundborg M, Karlén A
   *Is it possible to increase hit rates in structure-based virtual screening by pharmacophore filtering? An investigation of the advantages and pitfalls of post-filtering.* J. Mol. Graph. Mod. 2008, 26, 1237-1251

IV. Muthas D, Ericsson DJ, Lindeberg G, Unge T, Karlén A
   *Identification of novel non-peptidic inhibitors of Mycobacterium tuberculosis ribonucleotide reductase via a combined shape and structure based virtual screen.* Manuscript

Reprints of Paper I and II are made with kind permission from John Wiley & Sons, and Paper III is reprinted with kind permission from Elsevier Inc.
Contents

1 Introduction................................................................................................11
  1.1 Drug design ........................................................................................11
  1.2 Rational approaches to pathogen inhibitor discovery ........................12

2 Tuberculosis...............................................................................................13
  2.1 A short history of tuberculosis ...........................................................13
  2.2 Tuberculosis and Mycobacterium tuberculosis ..................................14
  2.3 Treatment of tuberculosis...................................................................15
  2.4 Ribonucleotide reductase ...................................................................16
    2.4.1 RNR inhibitors targeting subunit association .............................17
    2.4.2 Crystal structures of RNR...........................................................19

3 Molecular modeling ...................................................................................21
  3.1 Experimental design...........................................................................21
    3.1.1 Statistical molecular design ........................................................21
  3.2 Multivariate statistics .........................................................................22
  3.3 Molecular docking..............................................................................23
    3.3.1 Overview of docking methodology ............................................23
    3.3.2 Posing .........................................................................................23
    3.3.3 Scoring........................................................................................24
    3.3.4 Virtual screening.........................................................................25

4 Aims of the thesis.......................................................................................27

5 Development and application of a novel conservative design for fast
  followers .......................................................................................................29
  5.1 The design concept.............................................................................29
  5.2 Novel amino acid scales .....................................................................30
  5.3 FHDoE ..................................................................................................30
    5.3.1 Coverage of the experimental domain ........................................32
    5.3.2 QSAR modeling based on a FHDoE library...............................33
  5.4 Conclusion..........................................................................................34

6 Design and analysis of a peptide library targeting ribonucleotide reductase
......................................................................................................................37
  6.1 Investigated peptide libraries..............................................................37
  6.2 FHDoE library targeting Mtb RNR....................................................38
6.3 Conclusions ..................................................................................................41

7 Improvement of enrichment rates in virtual screening using pharmacophore filters .................................................................43
  7.1 Introduction ................................................................................................43
  7.2 Test sets and docking software ................................................................43
  7.3 Generation of pharmacophores ..............................................................45
  7.4 Evaluation of pharmacophore post-filtering .........................................45
  7.5 Results and discussion .........................................................................46
  7.6 Conclusion ..............................................................................................48
  7.7 Comparison to other methods ...............................................................48

8 Identification of non-peptidic inhibitors of ribonucleotide reductase ......51
  8.1 Shape-based screening ........................................................................51
  8.2 Structure-based screening ....................................................................53
  8.3 Biological evaluation ............................................................................53
  8.4 Conclusion .............................................................................................54
  8.5 Future outlook .......................................................................................54

9 Conclusions ..................................................................................................55

10 Acknowledgment .......................................................................................57

11 References ................................................................................................61
### Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADMET</td>
<td>Absorption, distribution, excretion, metabolism, and toxicity</td>
</tr>
<tr>
<td>CDK2</td>
<td>Cyclin-dependent kinase 2</td>
</tr>
<tr>
<td>COX2</td>
<td>Cyclooxygenase 2</td>
</tr>
<tr>
<td>DoE</td>
<td>Design of experiments</td>
</tr>
<tr>
<td>DOOD</td>
<td>D-optimal onion design</td>
</tr>
<tr>
<td>DOTS</td>
<td>Directly observed therapy, short-course</td>
</tr>
<tr>
<td>ERα</td>
<td>Estrogen receptor α</td>
</tr>
<tr>
<td>fXa</td>
<td>Factor Xa</td>
</tr>
<tr>
<td>FD</td>
<td>Factorial design</td>
</tr>
<tr>
<td>FFD</td>
<td>Fractional factorial design</td>
</tr>
<tr>
<td>FHDoe</td>
<td>Focused hierarchical design of experiments</td>
</tr>
<tr>
<td>HDoE</td>
<td>Hierarchical design of experiments</td>
</tr>
<tr>
<td>HSV</td>
<td>Herpes simplex virus</td>
</tr>
<tr>
<td>MAPS</td>
<td>Minimum analogue peptide sets</td>
</tr>
<tr>
<td>MDR TB</td>
<td>Multidrug resistant tuberculosis</td>
</tr>
<tr>
<td>MMP3</td>
<td>Matrix metalloprotease 3</td>
</tr>
<tr>
<td>Mtbe</td>
<td><em>Mycobacterium tuberculosis</em></td>
</tr>
<tr>
<td>NA</td>
<td>Neuraminidase</td>
</tr>
<tr>
<td>OPLS</td>
<td>Orthogonal PLS</td>
</tr>
<tr>
<td>OPLS-DA</td>
<td>Orthogonal PLS discriminant analysis</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal component analysis</td>
</tr>
<tr>
<td>PDB</td>
<td>Protein data bank</td>
</tr>
<tr>
<td>PLS</td>
<td>Partial least squares projection to latent structures</td>
</tr>
<tr>
<td>QSAR</td>
<td>Quantitative structure–activity relationship</td>
</tr>
<tr>
<td>RAPID</td>
<td>Rational approaches to pathogen inhibitor discovery</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver operating characteristics</td>
</tr>
<tr>
<td>RNR</td>
<td>Ribonucleotide reductase</td>
</tr>
<tr>
<td>SAR</td>
<td>Structure–activity relationship</td>
</tr>
<tr>
<td>SBVS</td>
<td>Structure-based virtual screening</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>SMD</td>
<td>Statistical molecular design</td>
</tr>
<tr>
<td>Sty</td>
<td><em>Salmonella typhimurium</em></td>
</tr>
<tr>
<td>TB</td>
<td>Tuberculosis</td>
</tr>
<tr>
<td>tciz</td>
<td>Theoretically derived and chemically intuitive z-scales</td>
</tr>
<tr>
<td>VS</td>
<td>Virtual screening</td>
</tr>
<tr>
<td>WHO</td>
<td>World Health Organization</td>
</tr>
<tr>
<td>XDR TB</td>
<td>Extensively drug resistant tuberculosis</td>
</tr>
</tbody>
</table>
1 Introduction

1.1 Drug design

Historically, serendipity and trial and error have played a major role in the discovery of drugs. The source of active substances has often been medical plants and herbs. With the advent of synthetic organic chemistry and modern pharmacology a more systematic search for new pharmaceutically active compounds began, and these were often evaluated using animal experiments.¹

Today, drug discovery usually follows the general scheme presented in Figure 1. In short, the goal is to identify a compound that can modulate the effect of a molecular target that regulates a biological process related to a disease. This target is often a receptor or an enzyme. G-protein-coupled receptors together with nuclear hormone receptors and ion-channels are the most exploited types of targets of the drugs currently on the market.² Once a target has been identified and shown to be relevant in a disease model, high-throughput screening, or its theoretical counterpart, virtual screening (VS), is usually employed to generate a set of hit compounds. The hits are analyzed and usually clustered into structural classes. Following this, some of the promising clusters that show good physico-chemical properties and also have a commercial potential are subject to further chemical exploration. The synthesized compounds in these lead series are evaluated and a structure–activity relationship (SAR) is derived, as well as the pharmacokinetic and pharmacodynamic profiles of the most promising compounds. The final phase in preclinical research is the transformation of a lead structure into a candidate drug, which is then considered for testing in clinical trials.

It is estimated that it usually takes about 10–15 years to develop a drug, and that it costs about US$ 800 million.³ So in order to decrease the risk of downstream attrition, many processes are run in parallel to assure that the hits generated can be transformed into good leads, which can in turn become good candidate drugs with a high probability of succeeding in clinical trials. Target validation is a continuous process and is often run throughout development. The modern drug discovery and development process is described by Bleicher et al.⁴ and well illustrated by the development of aliskiren,⁵ and the anti-influenza drugs zanamavir and oseltamivir.⁶
Computational modeling plays an important role in predicting e.g. the properties and activities of the various compounds at all stages of drug discovery.\textsuperscript{7, 8} This thesis is concerned with some of the computational techniques used in the early stages of the discovery process, and their application in anti-tubercular drug research.

1.2 Rational approaches to pathogen inhibitor discovery

The work presented in this thesis has been conducted within the RAPID (Rational Approaches to Pathogen Inhibitor Discovery) program at Uppsala University. RAPID involves interdisciplinary collaboration in which structural biologists, medicinal chemists and computational chemists are working together to identify inhibitors of novel targets for the infectious diseases tuberculosis (TB) and malaria. After a target has been identified, the structural biology unit starts the work of cloning, expressing and purifying the enzyme to enable assay development and crystallization studies. The medicinal chemistry group then tries to identify novel inhibitors by using virtual screening or design. Once hit structures have been identified, a synthetic chemistry program is started to establish a SAR, and an iterative process of modeling, synthesis, testing and crystallization of the compounds then follows.
2 Tuberculosis

2.1 A short history of tuberculosis

Tuberculosis is an ancient disease, manifested already in Pharaonic Egypt as shown by the examination of the bones of mummies. The disease has continued to plague mankind, and has harvested innumerable lives, with new epidemics following closely in the footsteps of war and poverty. Although the symptoms of coughing up blood, high fevers and weight loss were all too familiar, its origin was unknown until 1882, when Robert Koch isolated and grew the contagious bacterium, later named *Mycobacterium tuberculosis* (Mtb). Koch’s discovery, together with advances in the development of antibiotics and antibacterial medicines, made it possible to develop drugs to combat this previously untreatable disease. In 1943, streptomycin became the first chemical with a proven clinical effect, and in 1952 Salman Waksman was awarded the Nobel Prize for its discovery. It was found by screening huge amounts of soil specimens for anti-tubercular activity. The discovery of streptomycin was closely followed by that of another anti-TB drug, paraaminosalicylic acid. It was discovered by the Swedish chemist Jörgen Lehmann at The Sahlgrenska Hospital in Gothenburg, and its discovery was very different from that of streptomycin, in that only a few compounds was initially suggested and tested. Both these drugs had good effect in treating TB, however resistance to both drugs emerged quite early on, and in an effort to overcome this, the first combination therapy was introduced in 1949.

The most potent anti-TB drug to date, isoniazid, was discovered simultaneously in 1952 by three companies, Bayer, Squibb and Hoffman La Roche, and it was possible for the first time to effectively treat TB; optimism spread and it was believed that the disease was finally under control, and that it no longer posed a serious threat. A steady decline in the incidence of the disease was also seen in most Western countries.

However, it was too early to declare victory over the disease. The increase in HIV/AIDS together with emerging resistance, showed that TB yet again presented a real threat to humanity. The World Health Organization (WHO) estimated that as many as 1.7 million people died of tuberculosis in 2006, and that about one third of the world’s population is infected with Mtb. Especially sub-Saharan Africa shows high incidence of TB (Figure 2), closely linked to the ongoing HIV epidemic. Furthermore, resistance to all
available anti-TB drugs has been reported, both as single drug resistance and multiple drug resistance (MDR TB), and fear of extensively drug-resistant tuberculosis (XDR TB) is justified. The prevalence of MDR TB is quite high in many countries from the former Soviet union, with up to 22.3% of all cases being MDR TB. In many of these countries XDR TB is also a big problem, with XDR fractions ranging from 4% – 24% of all MDR cases. However, XDR is not restricted to the former Soviet union, and has been reported on all continents.

This emphasizes the importance of continuing research to identify novel treatments for this ancient disease.

2.2 Tuberculosis and *Mycobacterium tuberculosis*

As mentioned above, TB is caused by the pathogenic bacterium Mtb. The disease spreads through the inhalation of bacteria-containing droplets from individuals with an active infection. Once inhaled, the bacteria are attacked by the host’s immune system and are phagocytosed by macrophages. The macrophages aggregate and form granulomas, also known as tubercles. The bacteria can thereafter lie dormant within these tubercles for decades, and the patient shows no symptoms of TB. However, when the immune system is weakened, for example, by age, malnutrition or HIV infection, there is a possibility of reactivation of the infection and relapse. For individuals with a normal immune system the risk of developing the disease after infection is
only about 10%, but when the immune system is compromised the risk can increase to 50%.23

Mtb is an extremely slowly dividing bacterium, with an outer cell wall rich in peptidoglycans and mycolic acids.24 This waxy envelope makes Mtb difficult to treat with chemotherapy since passage through this membrane is difficult.

2.3 Treatment of tuberculosis

The WHO recommends that tuberculosis patients follow the DOTS regime (directly observed therapy, short-course), which comprises a four-drug cocktail of isoniazid, pyrazinamide, ethambutol and rifampicin (Figure 3). All four drugs are given for two months to quickly kill the infectious bacilli and reduce sputum loads.25 During this period, the symptoms are usually reduced and the patient becomes non-contagious. In order to fully cure the patient and to combat resistance, a 4-month continuation phase with two drugs is recommended. Isoniazid and rifampicin are usually prescribed for this phase, but ethambutol is sometimes used instead of rifampicin. After more than 60 years, streptomycin still finds use in the clinical setting, where it is used, for example, to treat relapsed cases.

![Chemical structures of anti-tubercular drugs](image)

**Figure 3.** The five first-line anti-tubercular drugs.

To date, less than 30 compounds against TB have been approved or are in late clinical trials.26 The most potent anti-TB drug, isoniazid, is an irreversible inhibitor that interferes with cell wall synthesis. It works by inhibiting inhA after activation to an isonicotinic acyl anion or radical.27
Isoniazid kills actively growing bacteria and quickly reduces sputum loads, making the patient more or less non-contagious within 2 weeks. Cell wall synthesis is also the target of ethambutol, whereas rifampicin targets RNA synthesis by inhibiting RNA polymerase activity. The final recommended compound for DOTS, pyrazinamide, disrupt the membrane potential, thereby killing the bacteria. Unfortunately, all anti-TB drugs have side effects e.g. hepatotoxicity, skin reactions, gastrointestinal problems and neurological disorders.

2.4 Ribonucleotide reductase

Ribonucleotide reductase (RNR) is a key enzyme in the de novo synthesis of DNA, and it catalyzes the reduction of 2′-ribonucleotides to the corresponding 2′-deoxyribonucleotides. It has been exploited as a target in cancer therapy and antiviral drug development. There are three major classes of RNRs, and Mtb RNR belongs to the oxygen-dependent RNR class Ib.

The bioactive form of Mtb RNR is a tetramer composed of two heterodimers, each consisting of a large R1 and a small R2 subunit. The active site is located in the R1 subunit together with an allosteric specificity site that regulates the activity of the enzyme. Reduction is carried out via a radical transfer, where the radical is generated on Tyr on R2, and transported via a network of residues to the active site in R1. This means that the two subunits have to associate for the enzyme to be active.

RNR contains several sites that can be targeted in order to inhibit RNR activity (Figure 4). Four RNR inhibitors are currently in use in the clinical situation: gemcitabine, fludarabine, clofarabine and hydroxyurea, of which the first three are antimetabolites targeting the active site, while hydroxyurea targets radical generation. Two compounds are currently undergoing clinical trials, the R2 down-regulating antisense therapeutic, LTO-2040, which is in phase II studies for cancer treatment, and the small interfering RNA, CALAA-01, which has just entered phase I testing.
Mtb has genes coding for two different R2 subunits and also for a class II RNR. The genes coding for the active Mtb RNR complex (with the large R1 subunit, nrdE, and the small R2-2 subunit, nrdF2) have been shown to be required for optimal bacterial growth in knockout studies and transposon hybridization studies. Furthermore, Yang et al. showed that Mtb RNR could be inhibited by an acetylated peptide corresponding to the 7 C-terminal residues (I, Figure 5) of the R2-2 unit with an IC$_{50}$ of 20 μM.

Figure 4. RNR exhibits many possible starting points for inhibitory design. The inhibitors discussed in this thesis are aimed at preventing subunit association.

Figure 5. The acetylated peptide derived from the 7 C-terminal residues of the biologically active R2-2 subunit of Mtb RNR.

2.4.1 RNR inhibitors targeting subunit association

The first study reporting that peptides corresponding to the R2 C-terminal inhibited RNR was presented in 1986 and concerned the herpes simplex virus (HSV). This approach to inhibition has since been shown to be applicable to other species, e.g. the mouse, and Mtb.
Scientists at Boehringer Ingelheim have performed a great deal of research on HSV RNR inhibitors, in which they transformed the C-terminal nonapeptide into a series of potent peptidomimetics (Figure 6). These peptidomimetics were shown to specifically inhibit the HSV RNR and to have an antiviral effect in both cell assays and in vivo models.

Extensive research has also been carried out on mammalian RNR subunit association inhibitors as potential anti-cancer drugs. The Cooperman lab has published several studies in which oligopeptides were employed to inhibit
mammalian RNR subunit association. They have reported both active turn mimetics and small peptides with good activity (Figure 6).  

2.4.2 Crystal structures of RNR

There is to date (16th Mar. 2009) no publicly available crystal structure of the large R1 subunit of Mtb. However, five structures of R1 from *Salmonella typhimurium* (Sty) have been reported in PDB\(^5\) (PDB ID: 1PEM, 1PEU, 1PEO, 1PEQ\(^6\) and 2BQ1\(^6\)). The R1 subunits of Mtb and Sty share about 71% sequence identity, with even larger conservation in the R2 binding site. The four 1PE\(^*\) structures (see above) contain the R1 subunit both in its native form and with different effectors.\(^6\) 2BQ1 (Figure 7) is a crystal structure of the Sty R1-R2 holoenzyme that also shows density for the bound, highly flexible R2 C-terminal. Since this binding region of R1 is highly conserved compared to Mtb it makes a good model system for Mtb. Unfortunately, the resolution is quite poor (4.0 Å) and only a few side chains of the C-terminal R2 are resolved (among them Trp\(^{317}\) and Phe\(^{319}\)). The small subunit of Mtb has been crystallized (PDB ID: 1UZR)\(^3\), however the C-terminal is unordered.

A crystal structure has also been reported for *E. coli* RNR R1, together with a 20-mer peptide corresponding to the R2 C-terminal,\(^6\) as well as structures of *S. cerevisiae* RNR R1 together with the corresponding R2 C-terminal heptapeptide and a peptidomimetic hexapeptide.\(^6\) These structures are, however, quite dissimilar to Mtb, showing less than 30% sequence identity.

![Figure 7](image). Structure of *Salmonella typhimurium* ribonucleotide reductase holo-complex (PDB ID: 2BQ1). The right image shows the binding of the R2 C-terminal into a hydrophobic cleft of R1.
3 Molecular modeling

Molecular modeling is used throughout the drug discovery and development process. It involves the systematic use of many different computational tools aimed at predicting a wide range of physico-chemical properties, target–ligand interactions, ADMET (Absorption, distribution, excretion, metabolism, and toxicity) properties, etc. The two modeling techniques used most in the work described in this thesis, statistical molecular design and virtual screening, are usually employed early in drug discovery to identify novel hit structures and to generate information-rich compound libraries for lead generation and optimization.

3.1 Experimental design

A great deal of thought and planning should be devoted to an experiment before it is performed. It should be clear why the experiment is being conducted, and what the goal of the experiment is. It must also be possible to interpret the results. Design of experiments (DoE) is a broad term used to describe different algorithms and methods used to decide how the experimental factors should be varied in order to extract as much information as possible in as few experiments as possible. In other words, a sufficient number of experiments, no more and no less, should be performed to address a specific question.

In order to maximize the utility of experimental design, time and effort should be devoted to understanding the system being modeled. After deciding which factors should be explored and which experimental domain is to be sampled, the most suitable design should be chosen. Once the design matrix has been generated it should be inspected to ensure that it has good quality, i.e. low condition number, high resolution, few confoundings, and small variance–covariance values.

3.1.1 Statistical molecular design

Statistical molecular design (SMD) is an extension of experimental design that generates information-rich chemical libraries. In order to generate these libraries, the chemicals are represented by numbers, via chemical descriptors such as e.g. molecular weight and CLOGP, or by
using experimentally derived values. The number of available descriptors is usually large, and it is common practice to use principal component analysis\textsuperscript{72} (PCA) (see below) to extract a few principal properties. These principal properties (or original variables if they are not compressed) are then treated as factors to be varied in a specific experimental design.

Two of the most straightforward experimental designs, besides trivial “random picking”, are the factorial design (FD) and the fractional factorial design (FFD).\textsuperscript{69} In these methods each factor is investigated independently of the others. Two-level (F)FD is often used, where each factor is explored at two settings, and the experimental domain will thus correspond to a square, box, or hyperbox in chemical space, defined by the high/low settings of the investigated factors.

Chemical space is, however, not always perfectly suited for (F)FD since it can be difficult to identify natural corner experiments, and in such cases other design strategies may be more suitable. One of the more popular designs is the D-optimal design.\textsuperscript{73} This algorithmic design identifies experiments that span the experimental domain as much as possible by maximizing the variance–covariance matrix of a certain regression model. This means that the experiments span as large a volume as possible. With a linear regression model, the D-optimal design does not cover the inner part of the experimental domain well, and the D-optimal onion design (DOOD) was developed to address this.\textsuperscript{70, 74, 75} This design methodology divides chemical space into layers, in a similar fashion to space-filling designs, which are then sampled individually to ensure better coverage of the inner regions of the experimental domain.

Before carrying out SMD, it must be decided whether the design should be made in building block or product space. It has however been shown that SMD can be performed in building block space without losing diversity in product space.\textsuperscript{76} This means that only a fraction of the possible compounds must be characterized, while this subset still contains most of the information present in the whole data set.

### 3.2 Multivariate statistics

PCA is a projection method that can be used to investigate how observations in large data sets are related to each other. It compresses the data into fewer, orthogonal principal components. Each point \((N)\) in a data matrix \(X\) \([N,K]\) is projected onto an \(A\)-dimensional \((A<K)\) hyperplane so that the maximum variance in \(X\) is described. Through this projection, \(X\) is transformed into a score matrix, \(T\), and a loading matrix, \(P\), according to the equation below. The scores describe how the observations are related to each other, whereas the loadings indicate how the original variables contribute to the principal components making up the hyperplane. The variation of \(X\) explained by
these two new matrices is expressed as an $R^2$ value, where values close to 1 indicate that most of the variation present in $X$ is captured by the projection. The residuals are collected into the matrix $E$.

$$X = TP'E = t_1p'_1 + t_2p'_2 + t_3p'_3 + ... + t_Ap'_A + E$$

The regression counterpart to PCA is partial least squares projection to latent structures (PLS), where the correlation between the score vectors of a variable matrix $X$ and a response matrix $Y$ is maximized. By doing this, both $X$ and $Y$ are approximated using latent variables and their relationship is modeled at the same time. The number of components that should be extracted is usually determined via cross-validation. A benefit with PLS over for example multiple linear regression is that PLS can handle correlated variables and missing observations. This makes PLS a good choice for regression analysis when dealing with “short and fat” matrices, such as those often generated in drug design.

A complicating factor in PLS is that all variation in $X$ is modeled to be related to $Y$. This is however not always the case since there is often variation in $X$ that is uncorrelated to $Y$. This means that the suggested PLS model might be hard to interpret due to the influence of this structured noise. In 2002, Trygg and Wold presented orthogonal PLS (OPLS), which is a method that removes variation in $X$ that is unrelated to the response $Y$. This procedure makes the interpretation of the OPLS model easier, since all the predictive power of the model is collected in the predictive components. The resulting orthogonal components should also be inspected to understand the uncorrelated variance in $X$.

3.3 Molecular docking
3.3.1 Overview of docking methodology
Docking involves the placement of a ligand within a binding site and the prediction of the free energy of binding for this pose. The goal is to find the global energy minimum of the complex, and numerous programs have been developed to solve this non-trivial problem. One of the first docking programs was DOCK, developed by Kuntz et al., which treated both the ligand and target as rigid bodies. Today, most algorithms deal with a flexible ligand and a rigid target, although significant effort has been devoted to also include protein flexibility, and the modeling of structural waters.

3.3.2 Posing
Docking of a ligand into a target structure can be divided into two steps, the first being posing. Posing is the task of identifying the most probable
conformation of a ligand within a binding site. The problem in posing is usually how the conformational space of the ligand and receptor should be covered in a sufficiently complete and at the same time fast way. Three main approaches have been applied to solve this problem: systematic searches, stochastic (random) searches, and simulation methods, for all of which several different approaches have been exploited.\textsuperscript{86}

In a systematic search one tries to sample all the available degrees of freedom, but it is usually not possible to explicitly examine all possible combinations due to the combinatorial explosion. One approach to reducing the search space is to “grow” the ligand incrementally inside the active site, and utilize a systematic search of the attached fragments. This approach is used, for example, in DOCK\textsuperscript{87} and FlexX.\textsuperscript{88} Another approach is to use libraries of pre-generated ligand conformations, as in FRED.\textsuperscript{89} This reduces the degrees of freedom to rotation and translation only. The docking program Glide\textsuperscript{90-92} attempts to sample the whole search space by first employing coarse filters, and when the number of possible conformations has been reduced, the sophistication of the search is increased. The posing engine of Glide also includes a stochastic search in the final steps of posing.

Another type of search method often utilized by docking programs is stochastic or random searching, in which the conformation of the ligands (and sometimes also the targets) is randomly changed. Monte Carlo and genetic algorithms are usually employed to generate the suggested changes. QXP\textsuperscript{93} (also known as Flo) employs Monte Carlo based searching of ligands within Cartesian space, whereas GOLD uses a genetic algorithm.\textsuperscript{94}

In the present work, Surflex\textsuperscript{95} was also employed as a docking engine. In Surflex each ligand is fragmented and the fragments are aligned to an idealized ligand called protomol. The best fitting poses in terms of protomol similarity and target complementarity are then selected.

3.3.3 Scoring

The second part of docking concerns the scoring of the suggested docking pose, i.e. the estimation of the change in free energy upon binding. It should be noted that if the posing algorithm has failed, any correct scoring is serendipitous. The scoring functions commonly used are usually divided into three distinct classes: force-field-based, empirical, and knowledge-based.\textsuperscript{96, 97}

Force-field-based scoring functions are based on physics and often score the target–ligand complex using a sum of the interaction energy and the internal energy of the bound ligand. Historically force-field based scoring functions calculated the pair wise atomic contacts between the ligand and the target, usually by calculating a Lennard Jones potential and an electrostatic Coulombic term. A common drawback of classical force-field-based scoring functions is that they usually have difficulties in modeling solvation and
entropic effects, and therefore better estimations of ligand electrostatics and solvation has been developed.

The most widely used type of scoring function are the empirical scoring functions, which are derived from fitting linear combinations of interaction terms, describing e.g. hydrogen bonds and hydrophobic contacts, to experimentally derived values. This makes them sensitive to the choice of training set, and any inherent variation in these experimentally determined values. Sotirrifer et al. published a study of the SFCscore, an empirical scoring function, based on more than 850 complexes, but even with this large training set the internal prediction errors in pKᵢ were in the range of one log unit, possibly indicating the limit of this type of scoring function.

The last category of scoring functions is the knowledge-based scoring functions. They have their origin in statistical physics, where distributions of geometries can be used to deduce an underlying potential. The target–ligand interaction is then modeled using these atomic interaction pair potentials and the goal is to implicitly explain the interactions governing binding. Since there is no parameter fitting to observed binding energies, it is less sensitive to choice of training set. However, there is not always sufficient data to calculate the distributions with certainty.

Since no single scoring function has been found to be superior, a common approach is to make use of several scoring functions and to derive a consensus score. This idea was first presented by Charifson et al. and is based on the assumption that using several scoring functions will make the combined score more stable. However, care should be taken to ascertain that the scoring functions used are not correlated, since this could lead to error propagation.

### 3.3.4 Virtual screening

The ultimate goal of VS is to enrich a hit list with true binders, and the method can be used to select a small set of compounds or to generate large libraries suitable for HTS. Virtual screening can be performed using either ligand or target information. Ligand-based approaches may employ similarity searching, pharmacophore matching or shape matching. In this work the shape-based virtual screening method ROCS from OpenEye Scientific Software was used. This method compares a set of pre-generated conformations to a search template molecule and scores them based on shape similarity. In structure-based VS (SBVS) molecular docking to a 3D structure of the target is performed.

Research concerning SBVS has grown enormously in the past decade. A search for reviews written between 2005 and 2009 concerning structure based virtual screening gave 74 hits in the ISI web of knowledge (16ᵗʰ Mar 2009). The reviews by Kitchen and Sousa give a good overview of the current state of the art. SBVS has also been evaluated in several
retrospective studies, but no particular program has consistently been found to be the best (see e.g. Kellenberger et al.\textsuperscript{109} and Kontoyianni et al.\textsuperscript{110}). Warren et al.\textsuperscript{111} showed how 10 docking programs and 37 scoring functions performed against 8 targets in the hands of program experts which, in my opinion, is one of the fairest comparisons of the performance of docking programs, given the fact that comparing programs is not a trivial task.\textsuperscript{112}

VS has emerged as an important tool in the early stages of drug discovery and is implemented throughout academia and industry. These methods have given rise to several interesting starting points, some of which have already lead to marketed drugs.\textsuperscript{113}
4 Aims of the thesis

The aim of the work described in this thesis was to develop new computational methods that can be used in early drug discovery and to apply them to the identification and design of novel, small-molecular compounds with anti-tubercular activity. The work was focused on Mtb RNR, and the transformation of the heptapeptide inhibitor 1 into a more drug-like small-molecule. To do this, new computational tools were developed, which were first evaluated using literature data, and then applied to RNR.

More specifically, the work involved the following steps.

- The generation of an SMD approach that biases a designed library towards a known lead structure in a rational way.

- The design of an information-rich peptide library based on the R2 C-terminal heptapeptide to probe the requirements for the inhibition of Mtb RNR.

- The investigation of the influence of pharmacophore post-filtering of virtual screening poses, to ascertain whether this is a viable approach to increase enrichment in cases with limited information.

- The identification of novel small-molecule inhibitors of Mtb RNR employing virtual screening.
5 Development and application of a novel conservative design for fast followers

In Papers I and II a novel design approach is presented and applied to Mtb RNR. In Paper I the method is described and compared to other design methodologies in terms of chemical space coverage and library quality. The strength of the method is that the designed libraries cover the chemical space surrounding the hit/lead structure in a good way, thereby increasing the probability of generating an informative and useful QSAR model. Paper II describes the design of a library of heptapeptides using this method. The library was synthesized and evaluated regarding biological activity, and an OPLS-DA model was derived that helped explain the requirements for binding.

The method was primarily derived for work with peptides, which often play important roles in biochemical processes, but can be easily extended to small-molecule design in building block space. Peptides are commonly used as starting points in medicinal chemistry, and understanding the specific interactions governing activity is pivotal for the efficient development of peptidomimetics to small-molecule drugs. The peptide–target interactions are often very specific, which means that activity is easily lost, e.g. a single amino acid exchange may be sufficient to render the peptide completely inactive.

5.1 The design concept

Statistical molecular design is often very useful since it helps to generate libraries that cover the available chemical space in a rational way. However, a potential problem when using SMD in large systems is that even small changes in each position can lead to a large total change. The method hierarchical design of experiments (HDoE) provides a means of remedying this issue. In this method, two design layers are combined to generate a library in which the “no-substitution” option is rationally incorporated. The method was applied to the design of melanocortin 4 receptor agonists, using a substituted phenyl as lead, allowing substitutions in up to five positions.

Paper I presents an extension of this idea where it is applied to the generation of peptide libraries. By defining a local allowed substitution
space for each amino acid position in the peptide, the resulting library will be even more biased toward the lead structure, while it still explores the space surrounding the lead. The combination of HDoE and focused building block selection produces the method called focused hierarchical design of experiments (FHDoE).

5.2 Novel amino acid scales

To be able to make a QSAR model starting from a set of peptides, the amino acids must be translated into numbers, something that has been the focus of several publications.\textsuperscript{116-126} The commonly applied z-scales of Hellberg et al. are the principal properties obtained from a PCA of both theoretical and experimental data on amino acids.\textsuperscript{117,118,123} Although very useful, the fact that experimental values are used in the characterization makes them hard to expand, since they require experimental testing of physical samples. Theoretical scales, such as those suggested by Norinder\textsuperscript{119} and Gottfries\textsuperscript{126} use several variables (both pure descriptors and principal properties) to describe each amino acid, which makes them easily extendable but more difficult to use in SMD since the number of compounds that must be synthesized to probe all factors becomes quite large. In this study, simple descriptors readily available within SYBYL\textsuperscript{127} were calculated and compressed using PCA to derive a theoretical amino acid scale (Figure 8), which can easily be expanded not only with new unnatural amino acids but also with hypothetical residues. The crude description of the residues keeps the number of components low. This theoretically derived and chemically intuitive z-scale (tciz) is suitable for designs, but is probably not the best choice for accurate QSAR modeling, without the addition of more molecular descriptors.

5.3 FHDoE

Designing libraries with FHDoE is a three-step process, in which the first two steps are the same as in HDoE, while the third step is a focused selection of building blocks (Figure 9).
In the first step of FHDoE each position is considered as a factor to be varied, and a FD/FFD is made to decide the substitution pattern. This design layer introduces the “no-substitution” setting in a rational way, considering both single and multiple changes in the native structure. This is followed by the generation of a substituent design matrix, which defines the principal property settings the new substituent should have. The two design matrices are required to have the same number of rows, which also corresponds to the number of designed peptides. These two designs are then “overlaid” to generate the final design matrix. The final building blocks are then selected from a local set of allowed substitutions. This means that the same principal property setting in two different positions does not have to correspond to the same part of chemical space (see Figure 9). Care should be taken to ensure that the final design is balanced and is of good quality.
Figure 9. A schematic of the FHDoE methodology. In the first design layer, the positions that are to be changed are determined. The second design layer determines how the changes are to be made. Building blocks are then selected from a local set of allowed substitutions. Note that the same principal setting in the final design matrix may give rise to very different substitutions, since each substitution is relative to the amino acid present in the hit/lead/native peptide.

5.3.1 Coverage of the experimental domain

The FHDoE method was compared to other common peptide design methodologies, such as the alanine and phenylalanine scan and minimum analog peptide sets. The experimental domain covered by the different methods was compared using a pentapeptide (Arg-Gly-Asp-Val-Tyr) model system. This anti-adhesive pentapeptide is derived from human leukocyte antigen DQ and has been mentioned as a potentially interesting anti-TB system. The peptide is challenging from a modeling perspective since it contains both glycine and arginine, which both have quite extreme properties (Figure 8). Glycine, for example, is the smallest amino acid, and is the only one with access to both D and L conformational space, making it important for the peptide’s secondary structure.

To investigate how well the different libraries covered the experimental domain, a PCA was performed of the 60 peptides that were designed with the different techniques. Each amino acid was described with the derived tciz scales. It can be seen from Figure 10 that no other method covers the
space around the native pentapeptide better than FHDoE. One possible reason is that glycine and arginine were retained in half of the peptides but still varied in a rational way. It can also be seen that the alanine and phenylalanine scans do not cover the space around the native peptide satisfactorily. Some of the peptides are very similar, while others are very dissimilar. This is not so surprising when one considers the investigated sequence, since replacing a valine with an alanine is quite a small change, whereas replacing tyrosine by an alanine represents a much larger modification. This again shows that these types of scans have to be interpreted with caution, and that several scans should be performed if a more complete picture of the binding site is required. It is clear from this analysis that the two focused SMD methods were superior regarding coverage of the relevant experimental domain.

Figure 10. Score plot from PCA of 60 pentapeptides derived using 5 different design methodologies. The native peptide is shown as a large dot. The FHDoE library is the only library that covers all corners surrounding the native structure. Note that global MAPS tries to screen the whole peptide space, without considering the native structure.

5.3.2 QSAR modeling based on a FHDoE library

Another important factor when considering the quality of the generated library is how it performs in QSAR modeling. To test the ability of FHDoE to generate an information-rich data set, a set of 58 dipeptides targeting angiotensin-converting enzyme, derived during the development of
Captopril, as previously reported by Hellberg et al., was used to assess the QSAR performance. A FHDoE training set consisting of 6 peptides was assembled using the peptide Val-Tyr as a hypothetical lead, describing the amino acids with two principal properties from the tciz scales. A two-component PLS model was then derived from these six peptides, and the remaining 52 peptides were predicted (Figure 11). The derived model was also compared to a model based on nine dipeptides selected using MAPS, and a model based on all 58 peptides. It was found that the FHDoE model provided a similar interpretation of activity as that by Hellberg et al. and that it also gave good predictions in the high-activity region, albeit using quite a small training set. The poorer predictions in the lower activity regions should not be a major problem once a high-activity compound has been identified. The fact that both the FHDoE and MAPS models showed better performance in the high-activity region than a model based on the whole data set, again illustrates the benefit of using an intelligent selection rather than “brute force”.

**Figure 11.** (A) The loading plot from the 2-component model derived from the 6-compound FHDoE training set. (B) Experimental vs. predicted activity of all 58 dipeptides, where it can be seen that the performance of the model deteriorates in the low-activity region, a fact that is counterbalanced by the good performance in the more interesting high-activity region. FHDoE training set shown as diagonal crosses.

### 5.4 Conclusion

Focused hierarchical design of experiments showed promise as a useful way of generating a library biased towards a known lead structure. It was demonstrated that the library generated probed the space close to the hit/lead/native structure well. However, the method suffers from the drawbacks that both design layers must comprise the same number of rows, and that there will always be some factor combinations that will not be explored. This can be partially overcome by including a few suitable experiments. The method also suggested a training set that made it possible to generate high-quality QSAR models for the high-activity region. The
method is also of a general nature and can easily be exploited in small-molecule design.
As mentioned above, RNR can be inhibited by peptides competing with the R2 subunit for association with R1. Yang and coworkers showed that a heptapeptide corresponding to the last seven C-terminal residues of R2-2 inhibited subunit association with an IC$_{50}$ of 100 μM in Mtb.$^{37}$ The acetylated heptapeptide (1, Figure 5) demonstrated even better potency, with an IC$_{50}$ of 20 μM. Although promising, this heptapeptide has several features that make it unsuitable as a drug: it has a rather low activity, it suffers from all the common ADMET-related problems inherent to peptides,$^{133}$ and it carries a net charge of -6. These issues must be resolved before interesting lead structures can be successfully derived for small-molecule design. The first step was to investigate the influence of the different side chains on activity.

6.1 Investigated peptide libraries

Three peptide libraries were evaluated in order to elucidate the requirements for binding (Paper II). The first library was aimed at identifying the minimum peptide length needed for activity, and contained a systematic truncation down to the tetrapeptide. It was found that all residues contributed to binding, and that truncation beyond the hexapeptide rendered totally inactive peptides (IC$_{50}>1$ mM). It was therefore decided that the remaining libraries would focus on the acetylated heptapeptide.

In the second library, a traditional alanine scan was performed. Here, each amino acid in 1 is replaced sequentially by an alanine. The basic idea behind an alanine scan is that it probes the relevance of each residue, since the possible interactions are minimized while the secondary structure of the peptide is conserved. Interestingly, all acidic residues could be replaced by an alanine without any marked drop in activity. This indicates that none of these residues is involved in any essential, specific electrostatic interaction.
Table 1. Summary of peptide libraries evaluated for Mtb RNR activity.

<table>
<thead>
<tr>
<th>Library</th>
<th># peptides</th>
<th># peptides with IC$_{50}$ &lt; 1 mM</th>
<th>Activity range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Native peptide</td>
<td>1</td>
<td>1</td>
<td>139 μM</td>
</tr>
<tr>
<td>Truncation</td>
<td>3</td>
<td>1</td>
<td>390 μM to &gt; 1 mM</td>
</tr>
<tr>
<td>Alanine scan</td>
<td>7</td>
<td>5</td>
<td>200 μM to &gt; 1 mM</td>
</tr>
<tr>
<td>FHDoE</td>
<td>16</td>
<td>5</td>
<td>170 μM to &gt; 1 mM</td>
</tr>
</tbody>
</table>

6.2 FHDoE library targeting Mtb RNR

As discussed above, the results from an alanine scan should always be treated with caution since it only probes how well replacing an amino acid with an alanine is tolerated. In order to obtain a more comprehensive conception of the peptide–R1 interaction, it was decided to employ the FHDoE strategy presented in Paper I. Although the alanine scan indicated that Trp$^5$ and Phe$^7$ were of great importance in 1, it was decided to replace these amino acids as well for the above-mentioned reasons. Since no single negative charge seemed to be essential in the peptide–R1 interaction based on the alanine scan, it was also decided to explore the influence of positive charge in all these positions. A library of 16 peptides was designed. Each position was described by two principal components from the tciz scales derived in Paper I, capturing the steric and polar tolerance of each position, yielding 14 factors. The allowed substitutions are shown in Figure 12.

Unfortunately, many of the designed peptides lacked activity. This was not unexpected since the starting peptide had poor activity, and it is more likely that activity will be lost than gained by replacements. However, a few interesting peptides were identified and some general conclusions could be drawn concerning the peptide–R1 interactions governing the association. It could be seen, for example, that Glu$^1$ of 1 could be replaced by a positively charged diaminobutyric acid without any marked decrease in activity, again indicating the relative unimportance of this residue, as also indicated by truncation and the alanine scan. Another interesting finding was that the tryptophan, identified as being essential in the alanine scan, could be replaced by both a naphthylalanine and benzothienylalanine, residues that are larger are more hydrophobic than the native Trp, and still exhibit activity. The most encouraging result was peptide 6 which had almost retained activity (IC$_{50}$ = 170±30 μM) but a reduced net-charge. Interestingly, peptide 7, in which all residues were replaced still showed some inhibitory effect (IC$_{50}$ = 790 ±170 μM).
Figure 12. Allowed substitutions in the FHDoE library. Note that only the side chains are drawn.
Since so many peptides lacked activity, and the range of activity was narrow, it was decided to perform a discriminant analysis to investigate if any of the positions were more important than the others regarding activity. All heptapeptides with a measured IC\textsubscript{50} better than 1 mM were defined as active, and the others were treated as inactive. An OPLS-DA\textsuperscript{77, 80, 134} was performed using the tciz variables together with a qualitative charge descriptor. The first analysis indicated non-linearity in the model and therefore cross-terms were added. Terms with small loadings in the predictive component were thereafter removed in an iterative fashion. A final two-component model based on 12 variables was derived explaining 75\% of the variance in X. The model had an internal predictivity of 45\% as judged by cross-validation (Figure 14).

Encouragingly, only one peptide was incorrectly classified (compound shown as large pyramid in Figure 14). The corresponding PLS-DA model was derived and Y randomization was performed further strengthening the model (R\textsuperscript{2} intercept = 0.396, Q\textsuperscript{2} intercept = -0.214). The data set was also divided into external test sets and new models were derived using the same descriptor set as in the original model. These new models were still robust, further indicating the validity of the original model.

Figure 13. Two of the most interesting peptides identified in Paper II.
Inspection of the loading plot (not shown) indicated that larger and more hydrophobic side chains are preferred in the Trp$^5$ position. A negative charge in position 3 also seemed to be important. This term may indicate that negative charge is important in order to guide the peptide to the binding site.

6.3 Conclusions

The fact that it was possible to derive a stable OPLS-DA model in this very challenging case demonstrated the utility of the method. The finding that larger more hydrophobic residues are tolerated in the Trp$^5$ binding site partly paved the way for the study presented in Paper IV. In conclusion, the FHDoE library contributed to the understanding of the peptide–R1 binding requirements, and also complemented the more classical approaches.
7 Improvement of enrichment rates in virtual screening using pharmacophore filters

7.1 Introduction

The posing algorithms of most docking programs are able to identify the bioactive conformation of the ligand under investigation, however, it is not always ranked as the best solution by the scoring function. One way to solve this problem is to visually inspect multiple poses of each compound. The most probable pose, based on information obtained from co-complexed structure(s) and other types of information, is then selected. However, this approach is not applicable in virtual screening where the number of suggested poses quickly increases far beyond the number that can be manually inspected. It would therefore be very helpful if one could use an automated filter to rule out erroneous poses. Since VS is often applied in the early stages of a drug discovery project, information about the requirements for binding is often limited or lacking. In these cases it would be beneficial to have a method that can exploit the limited information that is actually at hand. Paper III presents one such method, where a pose filter is created based on features from a single co-crystallized ligand. Furthermore, the influence of such a filter on the enrichment of structure-based virtual screening was also studied.

7.2 Test sets and docking software

Six different target proteins were used to evaluate the method, together with three different docking algorithms. The target proteins were selected so as to span a range of different binding sites, from small enclosed pockets to large open sites (Figure 15, Table 2). Only one crystal structure was used in docking for each target, and the specific complex was selected considering the size of the crystallized ligand and the resolution of the X-ray structure. The ligand data sets had been previously compiled and used by Jacobsson, and were comprised of active compounds from the literature together with approximately 10,000 decoy compounds complying with Lipinski’s rule of
To investigate the similarity of the decoys and the actives four simple descriptors were calculated and a two-component PCA model was derived. From this analysis it could be seen that the physico-chemical space of the different active classes were by and large covered by the decoy set. However, the neuraminidase inhibitors were smaller and more hydrophilic than the decoys in general, and some of the matrix metalloprotease 3 (MMP3) inhibitors had higher molecular weights than the average decoy.

Figure 15. Two of the six different active sites used in the study. The Connolly surface of the active sites is color-coded, showing hydrophobic residues in orange and hydrophilic residues in blue. (A) Neuraminidase has a highly polar binding site with many charged residues. (B) The active site of ERα is the other extreme in polarity, i.e. a very hydrophobic binding site with a few key hydrogen bonding groups at the far end of the pocket.

Table 2. Summary of targets used to evaluate the method (Paper III).

<table>
<thead>
<tr>
<th>Target</th>
<th>PDB code</th>
<th>Target type</th>
<th>SA$^a$ of co-crystallized ligand (Å$^2$)</th>
<th>Fraction PSA$^b$ of co-crystallized ligand</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDK2</td>
<td>1H00</td>
<td>Kinase</td>
<td>722</td>
<td>18%</td>
</tr>
<tr>
<td>COX2</td>
<td>6COX</td>
<td>Cyclo-oxygenase</td>
<td>599</td>
<td>25%</td>
</tr>
<tr>
<td>ERα</td>
<td>1L2I</td>
<td>Nuclear hormone receptor</td>
<td>548</td>
<td>20%</td>
</tr>
<tr>
<td>fXa</td>
<td>1IQE</td>
<td>Serine protease</td>
<td>857</td>
<td>23%</td>
</tr>
<tr>
<td>MMP3</td>
<td>1CIZ</td>
<td>Metallo-protease</td>
<td>780</td>
<td>19%</td>
</tr>
<tr>
<td>NA</td>
<td>1F8D</td>
<td>Viral coat glycoprotein</td>
<td>498</td>
<td>57%</td>
</tr>
</tbody>
</table>

$^a$SA – Surface area
$^b$PSA – Polar surface area
Three docking algorithms were used to avoid possible software bias: Zdock+, which uses a Monte Carlo based torsion driver, Surflex, which uses similarity measurements to an idealized ligand called protomol, and FRED, which employs rigid docking of pre-generated conformers.

7.3 Generation of pharmacophores

Peter Gund has proposed a widely accepted definition of a pharmacophore as a set of structural features responsible for a molecule’s biological activity. The present work was not concerned with deriving classical pharmacophores, but rather a set of potentially important features for binding (Figure 16). The features employed for filtering were identified automatically using the “Model query from alignment” option in UNITY. Only features located on the co-crystallized ligand were considered, and no information from the protein was used. The identified features were then grouped into two classes: a hydrophobic class (hydrophobes, aromatics) and a polar class (H-bond donors/acceptors, negative centers and positive N). The filters used for post-filtering were then created using different partial matching criteria and tolerance sizes (Table 3).

![Figure 16. Examples of identified features in the co-crystallized ligand of 1L2I. Green spheres correspond to hydrophobic/aromatic features and cyan spheres to polar features (here each cyan sphere corresponds to two features, a donor feature and an acceptor feature).](image)

7.4 Evaluation of pharmacophore post-filtering

The method was evaluated by comparing the receiver operating characteristics (ROC) curves and enrichment factors obtained using only the top scoring pose from docking to those after filtering. Each compound in the docking set had been passed through LigPrep from Schrödinger to generate different stereoisomers, tautomers and protonation states. This
expanded set was docked into the six target structures using the three
docking algorithms. Ten poses were collected for each compound, and
enrichments and ROC curves were first calculated using only the top scoring
pose of each compound. These poses were then rigidly filtered using the
pharmacophores generated, and the highest scoring passing pose was then
considered as a hit, and any compound failing to generate a passing pose was
discarded and considered inactive. This procedure led to a reduction in the
number of compounds for ranking, and a re-ranking of the remaining
compounds. The procedure is illustrated in Figure 17.

Table 3. The filters used, number of features, and % of actives/decoys fulfilling the
conditions of stringent and allowing filter with 1 Å tolerance in a flexible search.

<table>
<thead>
<tr>
<th>Target</th>
<th>Hydrophobic</th>
<th>Polar</th>
<th>Stringent</th>
<th>Allowing</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDK2</td>
<td>3, 2, 1</td>
<td>11, 4, 3</td>
<td>59% / 26%</td>
<td>99% / 87%</td>
</tr>
<tr>
<td>COX2</td>
<td>4, 2, 1</td>
<td>3, 1, 1</td>
<td>100% / 86%</td>
<td>100% / 95%</td>
</tr>
<tr>
<td>ERα</td>
<td>4, 2, 1</td>
<td>4, 2, 1</td>
<td>93% / 42%</td>
<td>96% / 99%</td>
</tr>
<tr>
<td>fXa</td>
<td>4, 2, 1</td>
<td>10, 4, 2</td>
<td>17% / 18%</td>
<td>100% / 97%</td>
</tr>
<tr>
<td>MMP3</td>
<td>5, 3, 2</td>
<td>9, 3, 2</td>
<td>92% / 18%</td>
<td>96% / 85%</td>
</tr>
<tr>
<td>NA</td>
<td>0, 0, 0</td>
<td>14, 5, 3</td>
<td>98% / 61%</td>
<td>100% / 95%</td>
</tr>
</tbody>
</table>

*The bold number is the total number of features included in the filter. The number in italics corresponds to the number of features to be matched in the stringent setting, and the third to the number of features to be matched with the allowing filter.

bThe stringent setting demanded the fulfillment of at least 33% of all suggested polar features and 50% of all hydrophobic features.

cThe allowing setting demanded the fulfillment of at least 20% of all suggested polar features and 25% of all hydrophobic features.

7.5 Results and discussion

It was seen that enrichment was usually improved after post-filtering. Both
the area under the ROC curve and the enrichment factors benefited from
post-filtering. The method had the greatest impact on targets where scoring
did not perform well, but where posing was successful. This was seen, for
example, in ERα and NA with FRED (Figure 18). In these cases many of the
high scoring decoys were removed by the filter, leading to enrichment early
in the database. Encouragingly, when scoring had worked well, post-filtering
usually did not reduce the effect, indicating that filtering could be used with
confidence.
Figure 17. Schematic presentation of the method applied. The compounds are first ranked according to their top-scoring pose (A). Thereafter 10 poses from each compound are rigidly filtered using a pharmacophore. Re-ranking is performed using the passing poses (B).

Figure 18. ROC curves for (A) ERα and (B) NA using FRED. It can be seen that the results are improved by using the suggested filters (dotted lines) compared to using the top-scoring pose (bold line).

Some lessons were learnt concerning the generation of the filters. The most demanding filter was probably too stringent, since too many compounds were often rejected, especially if the starting co-complexed compound
contained many potential interaction points. The most allowing filter had less effect while still improving the results. This implies that the filter should be trimmed so that the database is reduced to a size close to the desired number of compounds.

There are many potential problems in molecular docking, one being the inability to identify the bioactive conformation. This means that the scoring functions have a hard task to distinguish true binders from false positives. In this method it is required that the docked pose matches a set of predefined interaction points before being scored. This leads to the poses being placed in a similar frame of reference, making it easier for the scoring function to identify the true binders. This is exemplified in the case of ERα, where the performance of all the scoring functions was more or less the same after filtration, although they showed very different performance before filtering.

The method will probably be most valuable when only a limited number of compounds can be tested, i.e. when only a small number of false positives can be tolerated. The filter may incorrectly reject active compounds with an unexpected binding mode that do not match the interaction points. This is, however, not a major problem in the above setting where it is sufficient to find a few true actives. If, on the other hand, a VS campaign is performed in order to identify a large set of compounds suitable for HTS, false negatives will present a greater problem and this method may not be the best possible approach.

7.6 Conclusion

The current method showed great promise. In most cases enrichment could be improved by employing simple filters derived from the structure of the co-crystallized ligand. This was often because many high-scoring decoys could be removed by filtration, i.e. the number of false positives was reduced.

If more information about binding requirements is available, better filters can probably be created leading to an even better filtering effect.

7.7 Comparison to other methods

A great deal of effort is being devoted to increasing the performance of VS, for example, by trying to include a priori information. Perola et al. have shown that inspection of known co-complexes of a certain target class can be used to construct pharmacophores that are very useful in virtual screening. However, the method requires knowledge about the target class, which is not always available at the beginning of a VS campaign.
Other commonly employed techniques include e.g. constrained docking, in which certain interactions are required to be fulfilled in the posing step. The possibility to employ different types of constraints has been implemented in most commercial docking programs. The main benefit of constrained docking is of course that the poses generated fulfill some key requirements set by the modeler, and that the poses generated are thus believed to be relevant. This method also depends on the fact that certain key interactions can be decided a priori. However, if the wrong interaction points are identified the method breaks down.

The method presented in Paper III has a number of benefits compared to the methods mentioned above. The need for a priori information is very limited, since one co-crystallized ligand is sufficient. Secondly, there is no constraint on the docking, so the whole docking space is still considered and is available for the modeler to inspect, and the filters can be adjusted subsequent to docking if required.
Identification of non-peptidic inhibitors of ribonucleotide reductase

Paper II describes a peptide library targeted at Mtb RNR. This library gave insights into the requirements of the different side chains for inhibitory activity. As previously mentioned, peptides are not well suited as oral drugs due to their poor stability and low bioavailability. The goal is therefore often to transform a potent peptide inhibitor into a small-molecule with better pharmacokinetic properties and retained activity.

In an effort to identify such small-molecules targeting RNR virtual screening was performed (Paper IV). It had been found in a previous study that some fmoc-protected amino acids could inhibit RNR. The best compound identified was 8, with an $K_{D2}$ of 15 μM (Figure 19), whereas 9 had an $IC_{50}$ of about 50 μM in a fluorescence polarization assay. Since 1 had a $K_i$ of 8.9 μM in this assay, compounds 8 and 9 were considered good starting points for finding more drug-like small-molecules.

Figure 19. Two identified fmoc-protected amino acids with inhibitory effect towards Mtb RNR, as judged by a fluorescence polarization assay.

8.1 Shape-based screening

In Paper II it was found that larger and more hydrophobic groups were tolerated in the Trp$^5$ position of 1. The corresponding Trp side chain of R2 resides in a hydrophobic cavity adjacent to that occupied by Phe$^7$ (Figure 20). These binding sites contain amino acids known to be enriched in protein–protein interaction “hot spots”, indicating that this region might be important for the interaction.
Since 8 and 9 are more or less equipotent, it was hypothesized that the fmoc moiety of 8 and 9 binds to the same binding site of R1. Compound 8 can be nicely overlaid with the last three residues of the co-complexed R2 C-terminal using ROCS (Figure 21). In this conformation the fmoc residue fills the hydrophobic cavity and the indole of 8 mimics the pi-cation interaction made by Phe\textsuperscript{7} to Arg\textsuperscript{685} of R1 (Figure 20). This conformation of 8 was used as the template for shape-based virtual screening using ROCS from OpenEye Scientific software.\textsuperscript{108} A database was generated containing about 1.5 million screening compounds from commercial vendors. The compounds were ranked using the combo and Tversky(q) scores included in the software, and about 1,000 compounds were retained for further analysis.
8.2 Structure-based screening

The compounds selected by shape-based screening were inspected for binding site complementarity using structure-based docking. As mentioned previously, there is no publicly available crystal structure of Mtb R1, but Mtb R1 shares about 71% sequence identity with the crystallized Sty RNR, and all amino acids within 8 Å of the R2 C-terminal are conserved between the structures. Therefore, this structure was judged to be a good template for docking. To ensure that the compounds were present in the relevant tautomer and protonation state, all compounds were passed through LigPrep from Schrödinger.

The expanded set of ligands was docked into a rigid protein, retaining 10 poses of each compound to increase the probability of retaining the correct pose for each compound. These poses were then passed through a pharmacophore, in a similar fashion to that described in Paper III. The pharmacophore contained hydrophobic features corresponding to the Trp and Phe side chains (Figure 21). About 400 compounds passed the pharmacophore, and these were visually inspected.

From these, 24 were selected for biological testing. Care was taken to select compounds that are as drug-like as possible, e.g. aromatic nitro groups and large aromatic structures were removed. The solubility of the compounds in water was also considered, favoring ortho-substitutions, polar groups and heteroaromatics.

8.3 Biological evaluation

After the study described in Paper II had been carried out a new assay was developed. This new fluorescent polarization assay measures activity as a function of competition with a fluorescent peptide binding to R1. The fluorescent peptide had a $K_D$ of 2.4 μM and is the heptapeptide corresponding to the R2 C-terminal elongated by Gly-Ser-Gly and a fluorescent dansyl moiety. This assay has a much larger throughput than the earlier used activity assay. All compounds were measured at 0.01, 0.1, 1, 10 and 100 μM in triplicate, using 0.6 μM R1 and 0.1 μM fluorescent peptide and a sigmoidal curve was fitted to obtain an IC$_{50}$ value.

Of the 24 compounds initially screened several showed activity. A potential problem is the influence of the total fluorescence intensity of a compound on the measured inhibition. This relationship was monitored to avoid false positives. Some of the more interesting compounds (Figure 22) were examined more closely, determining the activity at 12 concentrations ranging between 0.01 and 1000 μM.
8.4 Conclusion

The VS approach employed worked very well in identifying interesting non-peptide inhibitors of Mtb RNR. It is interesting to note that small drug-like inhibitors could be found that target this protein–protein interaction, and that show inhibitory effects similar to those of the much larger heptapeptide.

8.5 Future outlook

The most interesting compounds will be tested in a secondary assay similar to that described in Paper II to validate them as true hits. Once confirmed, a hit expansion will be performed to cover the chemical space surrounding some of the most promising hits. This expansion will hopefully allow a SAR/QSAR to be derived that will guide future medicinal chemistry research. The compounds will also be tested on bacteria to evaluate their anti-tubercular activity.
9 Conclusions

This thesis presents two novel computational techniques that have been designed to aid early drug discovery in a setting with limited prior information. They have also been applied to improve our understanding of the requirements for binding of a peptide to its target, and to identify small-molecule inhibitors of the potential anti-TB target ribonucleotide reductase.

First an SMD approach was developed which covered the interesting experimental domain well and that generated libraries suitable for QSAR modeling. This approach was then applied to elucidate the importance of the different amino acids in the heptapeptide inhibitor 1. It could be concluded that large hydrophobic residues were beneficial in positions 5 and 7, and that all acidic amino acids could be replaced, however, not all at the same time. Secondly, a post-filtering technique for virtual screening was tested. This method utilizes the information present in one co-crystallized ligand to identify potential features beneficial for binding. These features were then used as filters of docked poses with encouraging results in six test cases.

Based on a protected amino acid hit structure, combined shape- and structure-based virtual screening was performed. A database containing 1.5 million commercial compounds was screened using ROCS and thereafter docked using GlideXP. The poses suggested by Glide were then filtered using the method developed, and 24 compounds were tested regarding activity. Several of these showed activity in a fluorescence polarization assay. The most promising compounds had molecular weights below 400 Da and were almost as potent as 1.

In conclusion, small-molecule inhibitors of Mtb RNR have been identified, and two useful computational techniques developed to aid early drug discovery.
10 Acknowledgment

This work would not have been possible without the help and support from many people. Research is always a team effort!

First of all I would like to express my sincere gratitude to my supervisor Prof. Anders Karlén. Thank you for all your support and never-ending enthusiasm. Your vast knowledge in computational and medicinal chemistry has been a true inspiration. I do not understand how you always keep such a positive attitude and always manages to find time where there is none...

My co-supervisor Prof Anders Hallberg, thank you for accepting me as a Ph.D student, and for providing an excellent scientific and social atmosphere at the department.

I am also indebted to Prof. Torbjörn Lundstedt for introducing and guiding me in the world of design of experiments and multivariate analysis. You always see possibilities and never problems. I am also very grateful for your continuous support throughout these years and the fun times at conferences.

Dr. Yogesh Sabnis, for all the great science we have talked and done, all the good laughs in the office and all the fun times out of BMC. It is great having you as a friend and colleague.

I would also like to thank all present and former members of the comp-chem group: Dr. Christian Sköld; for never being wrong (although not always in a rational way…) and always helpful and cheerful. Dr. Shane Peterson; for all help with Linux issues, linguistic and scripting discussion. Dr. Anna Ax; for bringing synthetic chemistry expertise into our corridor. Dr. Aparna Vema; for great work on the HTS datasets and for always, always being positive. Martin Lindh; for being a great office mate full of inspiring ideas. Good luck with the projects. Dr. Micael Jacobsson; for many interesting discussions on science and many other topics, and also for providing me with the VS datasets and giving insights into the “modeling in the real world”. Anneli Nordqvist; for being a fellow supporter of the DoE approaches and the great work in the GS project. Dr. Wes Schaal; for sharing your vast computational knowledge, your humor, and it is great that you never really left the group.
Prof Mats Larhed for always having a clever answer to any question, for knowing all the ins and outs of the academic world, and for making the department such a nice place to work.

The TB group, it has been really interesting taking part in this research and putting modeling to work in a real setting. A big thanks to Johanna Nurbo for daring to try the FHDoE approach and to Dr. Gunnar Lindeberg for swift handling of peptides. I would also like to thank Dr. Luke Odell for talking me into actually performing some synthesis, and then take the time to guide me through the whole gruesome process. I would also like to thank soon to be Dr. Lagerlund. It has been great having you as a friend and colleague to share the anguish of thesis writing, pondering the future, finalizing the GS paper, playing with the kids, and other outside BMC activities.

Thanks to my M.Sc students Magnus Lundborg and Martin Lindh. It was great working with you, good luck with your coming theses. Thanks also to my Sofosko student Hossein Ashouri and research student Avinash Punekar.

Thanks to Gunilla Eriksson for making sure that everything works smoothly at the department and for making the extra effort to take care of us. Sorin Srbu for excellent computer maintenance, for being a solid rock when Windows behaves like a stormy ocean (especially just before handing in this thesis for printing...).

Special thanks to all who read and came with constructive criticism on the thesis; Martin, Pelle, Christian, Aparna, and Olle. Any errors are due to my own last minute changes.

Thanks to alla@orgfarm, for fun times at parties, ski-trips, coffee breaks and for all the good science. Special thanks to Alle for taking such a good picture in such lousy conditions.

The beer club, for all the laughs over tasty and not so tasty beers. I’ll definitely stay away from all blue beer in the future. I am also indebted to the solid scientific lessons that one can learn from just “looking around you”.

My collaborators at the structural biology unit; Daniel Ericsson for his hard work with testing RNR compounds, especially during this Christmas and spring. Thank you for going out of your way to make sure that everything was done in time! Annette for struggling with setting up an RNR assay, Torsten for always being positive and sure of success! Nisse for all his work with and help in finalizing the GS paper. Prof. Alwyn Jones is acknowledged for his enthusiastic leadership of the RAPID project.
I am indebted to Ass. Prof. Henrik Ottosson for introducing me to computational chemistry, giving me the possibility to visit Japan and inspiring me to pursue a PhD.

The Uppland football scene, especially IFK Uppsala, Sunnersta AIF and FF Citykamraterna 05 for letting me keep the rest of my body in trim. For interesting trips to interesting parts of Uppland, and also for all the laughs on and off the pitch. I would also like to give a special thank you to the board of Edsbro IF, who in a very clear way showed that the football I played have always been very serious and not just for fun…

My friends outside for keeping me in contact with the real world. Especially I would like to thank my “brothers”, for all the nice memories we share, and the memories we will share in the future.

The Larsson and Larsson-Kollberg families for welcoming me into to your fantastic family, and for always being supportive and helpful.

My mother Inger for always believing in me, supporting me in my life-choices and being a great grandmother. My father Hans for teaching me the need for dedication and hard work.

Most of all I would like to dedicate this to Ella, Valter, and Jennie, who make everyday meaningful. Without your love and support this would never ever have happened.
11 References

9. Much has been written about tuberculosis, its sufferers, its social and economical impact, and the development of medicines to treat it. For interested readers I would like to recommend Frank Ryans “The Forgotten Plague”, which deals with the development of the anti-tubercular medicines, and “The white death – A history of tuberculosis” by Thomas Dormandy.
14. See introductory Times magazine quote


42. Yang, F. D.; Spanevello, R. A.; Celiker, I.; Hirschmann, R.; Rubin, H.; Cooperman, B. S. The carboxyl terminus heptapeptide of the r2 subunit of mammalian ribonucleotide reductase inhibits enzyme-activity and can be used to purify the r1 subunit. *Fems Letters* **1990**, *272*, 61-64.


71. CLOGP, v4.3 with fragment database 23; BioByte Corp: Claremont, USA.


108. ROCS, 2.3.1; OpenEye Scientific Software: Santa Fe, USA.


127. *SYBYL*, 6.9.2; Tripos Inc: St. Louis, USA.
137. Cyclin Dependent Kinase 2 (CDK2), Cyclooxygenase 2 (COX 2), Estrogen receptor α (ER α), factor Xa (fXa), Matrixmetalloprotease 3 (MMP3), Neuraminidase (NA)
140. *FRED*, OpenEye Scientific Software: Santa Fe, USA.
142. *UNITY*, implemented in Sybyl v.7.3; Tripos: St. Louis, USA.
144. *LigPrep*, Schrödinger LLC: New York, USA.
Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Pharmacy 91

Editor: The Dean of the Faculty of Pharmacy

A doctoral dissertation from the Faculty of Pharmacy, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Pharmacy. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Pharmacy”.)