Modeling of Flexible Side Chains for Protein-Ligand Docking

Dissertation

zur Erlangung des Grades des
Doktors der Naturwissenschaften
der Naturwissenschaftlich-Technischen Fakultäten
der Universität des Saarlandes
von

Christoph Hartmann
aus Bonn

Saarbrücken
Februar 2008

Dekan:	Prof. Dr. Joachim Weickert
Tag des Kolloquiums:	9. Juli 2008

Prüfungsausschuss:

Vorsitzender:	Prof. Dr. Thorsten Herfet
1. Berichterstatter:	Prof. Dr. Thomas Lengauer, Ph.D.
2. Berichterstatter:	Prof. Dr. Hans-Peter Lenhof
Wissenschaftlicher Beirat:	Dr. Iris Antes

Die vorliegende Arbeit wurde von Januar 2005 bis Februar 2008 unter der Leitung von Prof. Dr. Thomas Lengauer, Ph.D. am Max-Planck-Institut für Informatik in Saarbrücken angefertigt.

Abstract

This work comprises new approaches that are developed to support structure-based drug design in cases where side-chain conformations are uncertain, be it through flexibility or the devised modeling procedure. A knowledge-based scoring function ROTA is derived that can successfully identify correct rotamers and near-native ligand placements. ROTA is also able to reliably estimate the binding affinity of a protein-ligand complex, even if the conformations of one or both binding partners contain small errors. The side-chain prediction algorithm IRECS is developed for generating protein models that contain ensembles of rotamers for flexible side chains. IRECS is guided by ROTA and can accurately predict single and multiple side-chain conformations that represent the flexibility and conformational space of the respective side chains. IRECS is also able to include knowledge of side-chain conformations from a homologous protein used as a template directly in its optimization procedure. A modeling and docking pipeline is constructed that comprises IRECS, ROTA and the docking program FlexE. This pipeline is tested on 40 targets of the screening database DUD, where it is shown that the application of ROTA and IRECS can significantly increase the performance of screening experiments in cases in which side chains are flexible or were modeled.

Zusammenfassung

Diese Arbeit stellt neue Methoden vor, die die strukturbasierte Suche nach Wirkstoffen in solchen Fällen unterstützen soll, in denen Seitenkettenkonformationen durch Flexibilität der Seitenketten oder durch die verwendete Modellierungstechnik nicht sicher bestimmt werden können. Die Bewertungsfunktion ROTA wurde abgeleitet um richtige Rotamere und Ligandplazierungen zu erkennen. ROTA ist außerdem in der Lage die Bindungsaffinität eines Protein-Ligand-Komplexes zuverlässig zu bestimmen, auch wenn die Konformationen der Bindungspartner geringe Fehler aufweisen. Das Programm IRECS wurde entwickelt um Proteinmodelle zu erzeugen, die Ensembles von Rotameren für flexible Seitenketten enthalten. IRECS verwendet ROTA zur Bewertung von Proteinkonformationen und kann zuverlässig Ensembles von Rotameren bestimmen, die die Flexibilität und den konformellen Raum der jeweiligen Seitenketten repräsentieren. IRECS ist auch in der Lage zusätzliche Informationen über Seitenketten eines homologen Proteins, das der Modellierung als Vorlage diente, während seiner Optimierungsprozedur zu nutzen. IRECS, ROTA und das Dockingprogramm FlexE wurden zu einer Modellierungs- und Dockingpipeline vereinigt und auf den 40 Proteinen der Screening-Datenbank DUD getestet. Es konnte gezeigt werden, dass in Fällen mit flexiblen oder modellierten Seitenketten die Anwendung von ROTA und IRECS die Leistung von Screening-Experimenten deutlich steigern kann.

Danksagung

Ich möchte an dieser Stelle all denen herzlich danken, die mich bei der Anfertigung dieser Arbeit in der einen oder anderen Weise unterstützt haben. Mein Dank gilt zuerst Professor Thomas Lengauer, der mich Anfang 2005 in seine Arbeitsgruppe Computational Biology and Applied Algorithms aufgenommen hat und meine Promotion betreut hat. Gleichfalls möchte ich Professor Hans-Peter Lenhof für seine Bereitschaft danken diese Arbeit zu begutachten. Ich möchte allen Mitgliedern unserer Arbeitsgruppe danken für viele interessante Diskussionen und viele kleine und größere Hilfestellungen, die ich erfahren konnte. Im Besonderen möchte ich Iris Antes dafür danken, daß sie mich während meiner Promotion wissenschaftlich begleitet und beraten hat. Auch Andreas Steffen und Andreas Kämper danke ich, da sie mir mit vielen Nachhilfestunden in organischer Chemie geholfen und nicht mit konstruktiver Kritik und motivierenden Kommentaren gespart haben.

Ich möchte Christian Lemmen, Holger Claußen und Marcus Gastreich von der Firma BioSolveIT danken, denn sie haben mich mit der Chemieinformatik vertraut gemacht und waren bei der Suche nach einem geeigneten Thema für meine Promotion sehr hilfreich. Ich möchte mich und im Namen meiner Ehefrau Julia und unseres Sohnes Dominik bei Brian Shoichet und seiner Arbeitsgruppe für die erfahrene Gastfreundschaft während meines Forschungsaufenthalts an der University of California, San Francisco danken. Ebenso möchte ich John Irvin und seiner Familie sowie Modesto und Caren Tamez für die alltägliche Hilfe und Gesellschaft fernab der Heimat danken. Diese Forschungsreise wäre nicht möglich gewesen ohne die finanzielle Unterstützung des Deutschen Akademischen Austausch Dienstes, dem ich dafür danken möchte. Auch für die Unterstützung von Ruth Schneppen-Christmann in Verwaltungsangelegenheiten und Joachim Büch bei IT-Problemen möchte ich mich herzlich bedanken. Auch möchte ich André Altmann, Iris Antes, Matthias Dietzen, Andreas Kämper und Ingolf Sommer danken, die diese Arbeit vorab gelesen haben und mir wertvolle Ratschläge zur Korrektur geben konnten.

Abschließend möchte ich mich herzlich bei meiner ganzen Familie bedanken. An erster Stelle meiner Frau, die mich immer bedingungslos unterstützt hat. Unserem Sohn Dominik verdanke ich viele heitere Stunden, aus denen ich viel Kraft für meine Arbeit schöpfen konnte. Zuletzt danke ich meinen Eltern und Schwiegereltern, die mir mit Rat und Tat zur Seite standen.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Molecules can recognize each other by having complementary shapes and chemical interaction patterns. Nature devices this ability to associate special chemical functions to pairwise interactions of molecules in a cell: like a key fits only in its own lock, two molecules can only bind to each other if they have complementary chemical and geometrical properties. Using these mechanism molecules can perform metabolic tasks and participate in higher-level cellular processes. In 1894 Emil Fischer was the first to connect the key-lock principle to targeted molecular interaction [52]:

> "Um ein Bild zu gebrauchen, will ich sagen, dass Enzym und Glucosid wie Schloss und Schlüssel zueinander passen müssen, um die chemische Wirkung aufeinander ausüben zu können."

This principle is the foundation of structure-based drug design: if the geometrical and chemical structure of a protein is known, one can construct a chemical compound, a drug, which has complementary chemical and geometrical properties to the protein binding site like the natural substrate does [16]. When a drug binds to its target protein, it imitates the natural substrate like a pick lock imitates a key and can block the binding site of the protein for further interactions.

A prominent technique known as protein-ligand docking, short *docking* in the context of this work, can support the drug design process by modeling the conformation of a protein-ligand complex and predicting the binding affinity between both binding partners. Any small molecule that binds to a protein is considered as a ligand here. The two major computational issues of docking are sampling of the relevant conformations of the protein-ligand-complex and scoring the molecular interactions between ligand and protein for each putative conformation of the complex. The most relevant application of docking programs in pharmaceutical industry is to run structure-based virtual screening experiments. Such experiments are used to filter whole chemical libraries with thousands of compounds for a small set of putative protein (non-covalent) binders. Those compounds with highest predicted affinity to the target are used as lead structures in further steps of the drug design process. Current docking programs can dock single ligands within minutes to seconds, which enables running screening experiments within a few days on small computer clusters. This speed is achieved using fast geometric algorithms like geometric hashing or pose clustering that create a set of putative placements of the ligand into the active site of the protein. However, most of these fast algorithms can only be applied efficiently under the rigid-protein-assumption. This assumption requires the protein to provide a fixed geometric scaffold for the final complex conformation. The protein therefore must not change upon ligand binding, just as it was expressed through the key-lock principle. However, the key-lock principle was extended by Daniel Koshland in 1958 [115]. His induced-fit theory entails three postulates:

> '... (a) a precise orientation of catalytic groups is required for enzyme action; (b) the substrate may cause an appreciable change in the three dimensional relationship of the amino acids at the active site; and (c) the changes in protein structure caused by a substrate will bring the catalytic groups into the proper orientation for reaction, whereas a non-substrate will not.'

Such changes can be easily observed by comparing experimentally derived 3D models of protein-ligand complexes such that the same protein binds different ligands [16]. For many protein-ligand complexes it has been observed that they can only be constructed successfully, if the docking program takes the protein flexibility into account. However, the induced-fit theory is in clear conflict with the rigid-protein assumption, and this has two main consequences for docking programs that simulate induced-fit effects: the first consequence is reduced speed, since the conformational space of the protein must also be sampled. The other concern is that as few irrelevant conformations of the protein should be sampled as possible, since this would enable a docking program to easily adapt the protein conformation to any screened compound. This would interfere with postulate (c) of Koshland and result in many wrong false-positive predicted binders that hamper the identification of true binders.

1.1 Motivation

This thesis addresses the problems of (i) finding relevant protein conformations for modeling induced-fit effects and (ii) appropriately scoring the interactions of proteins and ligands during binding. This thesis also places a special focus on the side chains of a protein interacting with the ligand. Side chains are much more flexible than the protein backbone and – as postulated by Koshland – the side chains of the active site are usually able to rotate and arrange themselves toward the ligand, like the pins of a modern cylinder lock can move to recognize a key. Docking methods can often ignore this issue if a protein model is provided such that the conformation of a binding ligand is already imprinted in the conformations of the side chains. However, such docking methods are especially prone to conformational errors in protein models. An analysis of the protein models submitted to a recent Critical Assessment of Structure Prediction (CASP) contest [149] revealed that a high ratio of side chains in the active site was assigned wrong conformations. This leads to the prediction of false-positive protein-ligand contact interfaces [38], which in turn renders such models almost useless for docking approaches that do not take side-chain flexibility into account. This thesis therefore aims to answer the following questions:

1. How can side-chain flexibility be predicted and modeled?
2. How can correct and wrong conformations of a side chain be distinguished?
3. How can side-chain flexibility be included during docking so that a high number of protein binders can be identified with virtual screening?

The limitation to side-chain flexibility was chosen right from the start to simplify computational issues. There exists a number of proteins for which such a limitation during modeling and docking is admissible, but many proteins also require the inclusion of backbone flexibility. This issue is therefore discussed in the context of the methods presented in this work and – if reasonable – extensions are proposed that allow for taking both side-chain and backbone flexibility into account.

1.2 Overview

Chapter 2 gives a short introduction to the relevant topics of structural and computational biology and lists evaluation measures that are used throughout this work. An overview is then given on scoring of protein-ligand interactions, modeling of protein conformations and docking. Chapter 3 presents the ROTA scoring function together with the applied derivation technique and the results of a comparative evaluation. Chapter 4 addresses the first two questions of the previous section and describes the side-chain prediction program IRECS that can also handle flexible side chains. Chapter 5 comprises an extension of IRECS that is meant to increase the accuracy of IRECS to cases in which the protein backbone was modeled using a similar protein as structural template. In Chapter 6 IRECS and ROTA are combined with FlexE to a complete modeling and docking pipeline that addresses the third question of the previous section. This pipeline is then tested and optimized using a large set of screening experiments. The sequence of these chapters follows the chronology of different steps taken during the project with the exception that the ROTA potentials for docking presented in Chapter 3 were derived previously of the screening experiments that are described in Chapter 6. The final discussion in Chapter 7 aims to identify application scenarios for which the presented methods appear to be most useful and discusses inconsistencies in data preparation and possible future extensions. Chapter 8 summarizes the achievements in methodology and results obtained in this work.

Chapter 2

Basic Techniques and Related Work

This chapter summarizes the tasks, materials and basic computational techniques that are relevant for structure-based drug design in general and for this work in particular. Most tasks have been known for long and so there are plenty of methods that try to give solutions. Thus, a selection of these approaches is presented here which is based on the impact of the selected approaches on the field, but also on the ability to provide a reference for comparison with methods developed in this work.

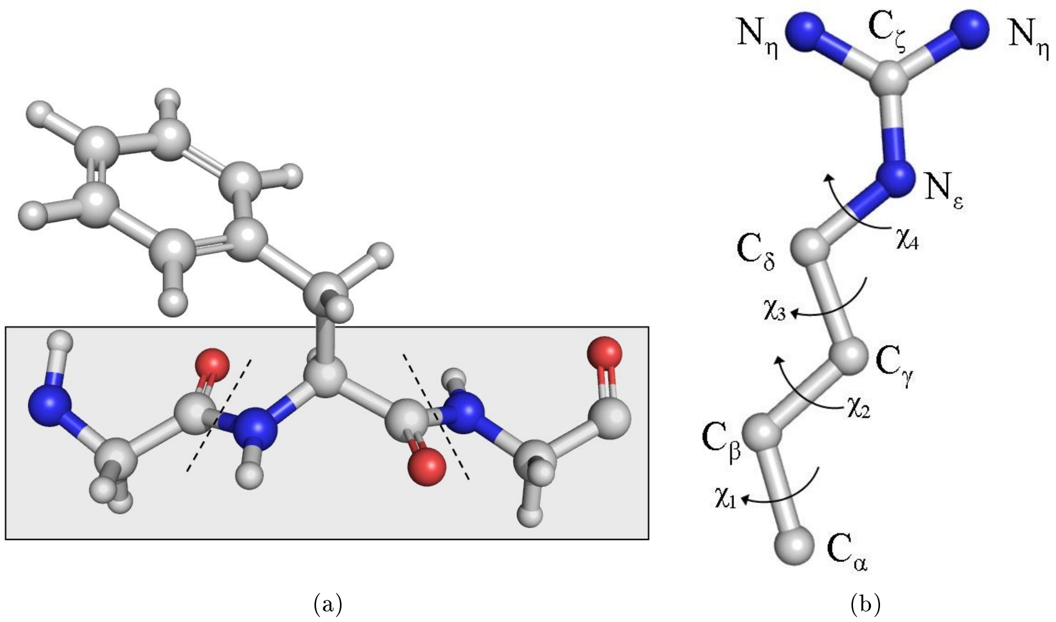
2.1 Relevant Topics in Structural Biology

A short review is given here that covers topics in structural biology that are most relevant for this work. A more detailed overview is given by Brandon and Tooze [22]. The following section describes the molecular entities of this work and their most important properties.

2.1.1 Molecular Entities

Proteins

Proteins are vitally important for cells – they participate in nearly every cellular process. From the chemist's point of view proteins are polypeptides, which consist of a sequence of amino acids that are assembled into long chains according to the genetic code of a given RNA sequence. The genetic code allows assembling proteins from 20 different standard amino acid types (although exceptions allow for more). Amino acids consist of a constant backbone part and a variable side-chain part, which can have different chemical properties, e.g. polar or apolar. By these properties the sequence of amino acids in the chain determines the structure, dynamics and functions of the overall protein (see Figure 2.1(a) for an example of a small peptide). Proteins can assume multiple spatial conformations, in which the side chains are usually much more flexible than the protein backbone. With the exception of proline, side chains are connected only through a single covalent bond between the C_β and C_α (see Figure 2.1(b)) to the remaining protein and are thus free to rotate. The flexibility of most side chains is therefore mainly limited by non-covalent interactions with their environment. The different conformations that a side chain is likely to adopt are called *rotamers*, the short form of 'rotameric isomers'. Rotamers are formed by internal rotations



(a)

(b)

Figure 2.1: Building blocks of a protein. (a) Glycine-phenylalanine-glycine peptide, the grey area highlights the backbone part. (b) Side chain of arginine.

represented by certain values of the dihedral angles (χ_1-χ_4, see Figure 2.1(b)) of a side chain. Proteins adopt various shapes to fulfill their manifold functions, e.g. as channels in the cell membrane, molecular motors, scissors, vises or transporters. They usually consist of many thousand atoms, which complicates the computation of pairwise interactions between those atoms. As a working assumption for docking programs, the relevant proportion of a protein for binding a ligand can be reduced to the so-called *binding pocket* or *active site*.

Ligands

All small molecules that bind to proteins – small natural substrates and drug candidates – are considered as ligands here. A molecule is small if it contains only about ten to twenty heavy atoms (all non-hydrogen atoms are usually considered as heavy atoms). Drug candidates are small molecules that bind to a target protein and fulfill certain requirements that concern their toxicity, absorption, metabolism, distribution in the body and excretion. Lipinski and coworkers [132] introduced a rule – the so-called rule of five – which sets constraints on the number of hydrogen-bond donors and acceptors, the molecular weight and the lipophilicity of a compound to be drug-like. This definition also includes small peptides (less than ten amino acids). Ligands can be highly flexible, and then sampling their many conformations requires much computational effort. Compared to proteins that are made up of twenty well-understood building blocks, the chemical space of ligands is more diverse [131]. This complicates the definition of a comprehensive scheme for scoring the non-covalent interactions which ligands can establish.

Cofactors are a special class of ligands and are often present in the active site where they

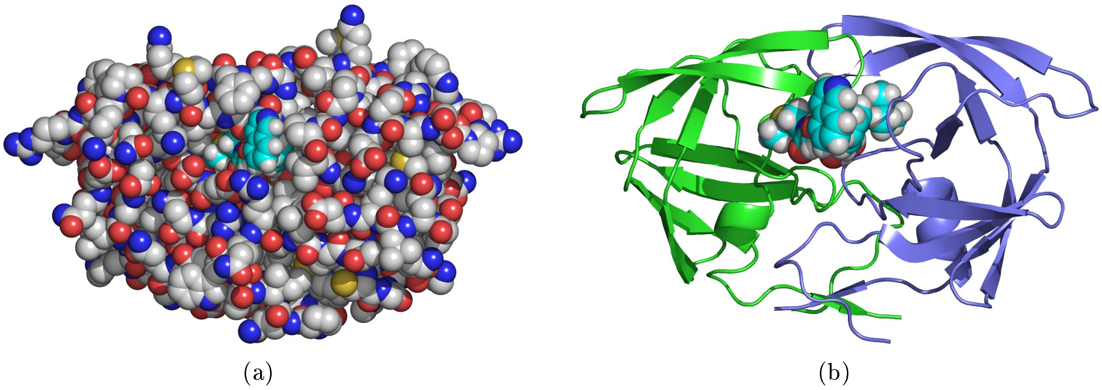


Figure 2.2: Different visualization of HIV protease (PDB ID: 1hpx). (a) All atoms displayed with spheres that represent their van der Waals radius (white=hydrogen, grey=carbon, blue=nitrogen, red=oxygen, yellow=sulfur). Hydrogen atoms are only shown for the ligand. (b) Secondary structure elements (helices, sheets and loops) of HIV protease are displayed. HIV protease is a dimer (green and blue chains), the flaps on top of the ligand are flexible and ment to 'cut' a peptide strand like scissors.

catalyze chemical reactions. Cofactors can also be single metal ions, which carry charges that enable them to exert strong attractive forces upon polar parts of ligands. These properties designate cofactors as being frequent key players in forming special protein-ligand complexes.

Solvent

All the previously introduced molecules are surrounded by water molecules in a cell. Water molecules can solvate other polar molecules by creating dipole interactions with them, the strongest being a hydrogen bond. The polarity of molecular groups and their ability to form hydrogen bonds determines whether they are hydrophilic (attract water) or hydrophobic (repel water). Two important processes that are involved in forming protein-ligand complexes are (i) establishing one or more hydrogen bonds between ligand and protein (which results in a partial desolvation of the ligand and the active site) and (ii) orienting hydrophobic parts of the ligand towards hydrophobic surface patches of the protein to keep water away from these areas, usually enabling the removed water molecules to create additional hydrogen bonds in their new environment. Many issues hinder the explicit handling of water molecules in time-critical modeling approaches. Such issues include their number, their mobility and their ability to reorientate easily. For that reason, water molecules are not treated as explicit entities in many modeling studies [124] as well as in this work, but solvent effects are included implicitly into the model of protein-ligand interactions.

2.1.2 Protein-Ligand Complexes

A protein offers binding pockets to its ligands that are complementary to the respective ligands in both their shapes and chemical properties. With respect to the complementarity of shape, the repulsive forces between atoms are most important, as they determine the

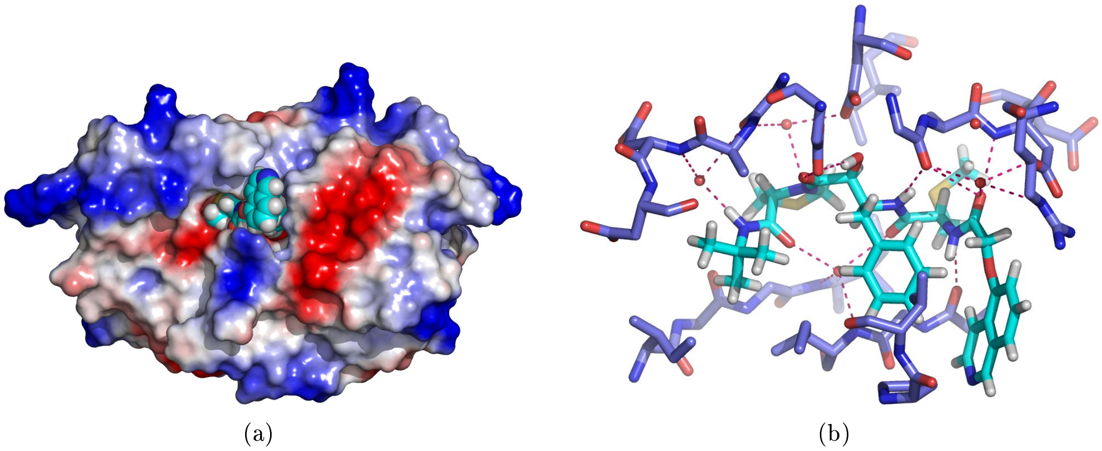


Figure 2.3: Complex of HIV protease with the inhibitor KNI-272 (kynostatin). (a) The surface of the protein is colored by preferences for electrostatics interactions. Charges are computed with the Amber 99 [221] force field and projected on the protein surface, whereas colored patches (red=positive, blue=negative) denote polar regions and white patches apolar protein regions. (b) The active site is displayed with ligand, water molecules (only oxygen atoms) and protein residues that interact with the ligand. Strong polar interactions are marked with dotted lines.

space taken by each atom and which space remains available for the ligand atoms. Among the electrostatic interactions that hold protein and ligand together, dipole interactions or interactions between complementary charged parts of the molecules [53] are all short-ranged interactions. Also, non-polar surfaces of proteins and ligands can attract each other in solution: since water molecules located at these hydrophobic surfaces are limited in their ability to establish hydrogen bonds, the strength of the binding can increase as the water molecules are pushed away by the ligand from these areas (this process is called *desolvation*), allowing them to build more hydrogen bonds with the surrounding solvent.

Usually, its flexibility enables the ligand to take on a conformation that complements the shape of the binding pocket and realize the interactions with the protein such as to minimize the free energy of binding. The protein itself is less flexible but, as stated in the introduction, the so-called induced-fit effects facilitate rearranging terminal hydrogen atoms, side chains or flexible loops to maximize complementarity of shape and interaction patterns, at the cost of spending energy to implement these conformational changes. Additionally, entropic effects play a role as the ligand looses degrees of freedom for translation, rotation and internal rotations upon binding (as does the protein), whereas the released water molecules gain degrees of freedom.

The total strength of binding or *binding affinity* is measured by the binding free energy ΔG_{bind} which is the difference between the free energy of the molecular system before binding and the free energy of it after binding:

$$\Delta G_{\text{bind}} = G_{\text{bound}} - G_{\text{unbound}} \tag{2.1}$$

ΔG_{bind} can be determined experimentally by measuring of the inhibition constant k_i which is the ratio of the on-rate k_{on} and the off-rate k_{off} of the binding process [24] (R is the gas constant and T the temperature):

$$\Delta G_{\text{bind}} = -RT \ \ln \left(\frac{k_{\text{on}}}{k_{\text{off}}} \right) = -RT \ \ln \ k_i \tag{2.2}$$

2.1.3 Protein Structures and Models

The central prerequisite for a successful docking run is a 3D model of the target protein that has sufficient quality. Such models can either be derived from experimental data or can be generated with computational methods. However, the accuracy of atom coordinates is a crucial issue for a docking run to succeed. If the model is not accurate enough, especially with respect to the position of atoms in the active site, the docking problem becomes a much harder task and special methods must be applied to increase the chance for a successful docking. Interestingly, these methods are closely related to those which are used for taking protein flexibility into account, which are discussed in Section 2.4.2. In the literature, it is common to use the term *structure* for a 3D model of a protein that is derived directly from experimental data. In most cases these are data from X-ray crystallography. Usually these models are the best source of information that one can get for the respective proteins, and therefore it is common sense to identify this model with the native conformation of the protein if the experimental method is known to produce highly accurate models. In contrast to this, the term *model* is usually applied to a protein structure generated by computer methods without experimental structure data on that protein. Although both terms actually denote a 3D model of the protein atom coordinates, this work sticks to this formalism for reasons of consistency.

X-Ray Structures from the Protein Data Bank

To date, X-ray crystallography provides 3D models of protein structures suited best for docking. The Protein Data Bank (PDB) [10, 11] is the largest source for publicly available protein structures solved by X-ray crystallography. Starting in 1972 with only a few deposited structures, the PDB contained more than 40,000 structures of proteins in June 2007. Most of the structures were resolved using X-ray crystallography (34,835 structures in June, 86.1%). The remaining structures were either resolved by nuclear magnetic resonance spectroscopy (NMR) (5,424 total, 13.4%), cryo-electron microscopy (EM) (101 total, 0.2%) or other methods (81 total, 0.2%). Although X-ray structures provide high-accuracy data of protein atom coordinates, these data do not provide perfect models [2]. The accuracy of coordinates of different atoms of a protein in a 3D structure depends on many factors like the type of protein or the refinement method. Given average Debye-Waller factors (also: temperature or B factors, a definition is given by Rhodes [176]) usually ranging from 5 Å^2 to 60 Å^2 (taken from the WHAT CHECK [82] server documentation[1]) for buried atoms, displacement of individual atoms through vibration in crystallized proteins is assumed to

[1] http://swift.cmbi.ru.nl/gv/pdbreport/checkhelp/explain.html#bfac

be within 0.25 Å and 1.0 Å (RMS displacement of different B-factors [175]: B factor of 5 $\equiv$ 0.25 Å, 20 $\equiv$ 0.5 Å , 79 $\equiv$ 1.0 Å).

There are also a number of errors in X-ray structures that result from insufficient experimental data (electron density maps) or unresolved discrepancies between the generated model and the experimental data [34, 108]. Such errors concern the conformations of side chains (especially histidine, asparagine, glutamine [226, 231] and surface residues), the identity of non-protein atoms, the position of solvent molecules, tautomeric states of ligands (since hydrogen atoms are usually not resolved), falsely connected secondary structure elements, and others. Moreover, crystal packing effects influence the conformations of protein-protein contact surfaces and side-chain conformations [29, 92].

Theoretical Models

Whenever a protein model is based on data different from first-hand experimental data it is classified as a so-called theoretical model, which implies that this model still needs to be confirmed by experiments. There exist many reasons that hamper scientists in generating the required experimental data, e.g. unavailability of proper experimental methods for a given target protein, wet lab capacity or firm time constraints. In such cases a number of techniques are available that can help in obtaining a protein model, e.g. homology modeling, which is described later in more detail (see Section 2.3). Usually, theoretical models are less complete, accurate and reliable than those models that are based on experimental data, but depending on the applied techniques, the skill of the modeler and the secondary knowledge on the modeling target, it is sometimes possible to generate protein models with sufficient quality for successful structural analysis [7].

2.2 Scoring Functions for Molecular Modeling and Docking

Scoring functions facilitate efficient computation of an estimate of the free energy of a molecular system. The exact calculation of the energy of a molecular system would require solving the Schrödinger wave equation for that system. Since this is computationally infeasible for large systems, a number of assumptions is made to enable the efficient computation of an approximate value (adapted from Schlick [188]).

- Molecular representation: the molecular system is regarded as a mechanical system. Atoms are represented as single bodies with a single center of mass. Molecules are considered as point masses such that atoms are connected by strong covalent bonds. Electrons are not represented explicitly but their interactions are represented by potential functions that apply to atoms and bonds. This assumption allows for using classical mechanics instead of quantum mechanics.

- Thermodynamic hypothesis: molecules assume configurations that minimize the free energy of the system. If the molecules in a system collectively reach such a configuration the system is said to be in thermodynamic equilibrium. This especially means that the observed (also: native) conformation of a protein is that of minimum energy.

This assumption is required to guide molecular modeling by evaluating the energy of a given configuration of the molecular system. This assumption requires the minimum energy state of the system to be unique and also stable (which is not generally true), meaning that small changes to the system (away from the thermodynamic equilibrium) do not cause large changes in the configurations of the molecules.

- Additivity: the energy of the system can be approximated by summing over the separate energetic contributions of interactions that are determined by simple structural features (like distances between atoms or dihedral angles). This assumption ignores more complex multiparticle terms in the energy function.
- Transferability: interaction potentials can be derived based on a representative set of structures of molecules and can be applied for predicting the structures of molecules for which they are representative.

Although there exist many approaches for computing the free energy with high accuracy [62], such approaches are rarely used in molecular modeling of large systems such as protein-ligand complexes due to prohibitive demands on computing time. In developing scoring functions one therefore usually aims at finding a compromise between accuracy and time-efficiency.

2.2.1 Potentials of Mean Force

This section reproduces the theory and formalism of so-called *potentials of mean force* (PMFs) that are the basis for the design and derivation of the ROTA scoring function which is introduced in Chapter 3. There are quite a number of scoring functions that apply PMFs to calculating the strength of pairwise interactions of atoms in a molecular system in dependency on their distance. The main advantages of PMFs are (i) that they can be derived based on structural data only, without the need for associated activity data, (ii) that no *a priori* knowledge about molecular interactions in the respective molecular system is required and (iii) that interactions with solvent are treated implicitly. The concept of PMF was founded by Kirkwood in 1935 [105] and later also formulated by Sippl [196]. The equations and declarations presented in this section mainly reproduce the work of Sippl, but variable names are changed such as to be consistent with the derivation of ROTA in Chapter 3.

Given a molecular system with a discrete set of states G. The Boltzmann law permits calculating for each state S of G the probability that the system will adopt this state, given that the energies of all states are known (k is the Boltzmann constant and T is the average system temperature).

$$P(S) = \frac{1}{Z} e^{\frac{-E(S)}{kT}} \tag{2.3}$$

Z is called the *partition function*.

$$Z = \sum_{\hat{S} \in G} e^{\frac{-E(\hat{S})}{kT}} \tag{2.4}$$

The central idea is that this law is invertible: if it is possible to determine the probabilities of the states, then the energy of these states can be calculated from the probabilities of the states by using this formula:

$$E(S) = -kT\left[\ln P(S))\right] - kT\left[\ln Z\right] \tag{2.5}$$

The rightmost term is constant , i.e. independent of the state of the molecular system. Given a molecular state of such a system in Cartesian space, one can calculate all pairwise Euclidean distances of atoms and store them in a matrix M. The conformation of all molecules in the system is determined by these distances. Using the assumption of additivity the energy of the system can be calculated by a sum of separable contributions from basic structural features. Sippl uses only pairwise atom distances between atoms a and b and ignores contributions by other structural features, i. e. bond angles or dihedral angles. The energy of the system can be calculated as a sum over all distances d in M between pairs of atoms that are part of particular chemical groups (also: *atom types*) a and b:

$$E(S) = \sum_{d \in M} E_{ab}(d) \tag{2.6}$$

The inverse Boltzmann law is now applied to the individual pairwise atom distances d and the probability $P_{ab}(d)$ of observing atoms of type and a and b at such a distance:

$$E_{ab}(d) = -kT\left[\ln P_{ab}(d)\right] - kT\left[\ln Z_{ab}\right] \tag{2.7}$$

As formulated by Sippl [197] $P_{ab}(d)$ can be approximated over a continuous sequence of distance intervals by counting the frequency of atoms of types a and b observed at a distance interval of size x in a representative dataset (x is usually set to values between 0.1 and 1.0 Å and named the *bin size*).

$$\lim_{n \to \infty} F_{ab}\left(\left[d - \frac{x}{2}, d + \frac{x}{2}\right[\right) \equiv P_{ab}(d) \tag{2.8}$$

However, the partition function Z cannot be determined with the same technique and therefore the energy can only be computed up to the constant terms $-kT\left[\ln Z\right]$ in Equation 2.5 and $-kT\left[\ln Z_{ab}\right]$ in Equation 2.7, respectively.

The computation of E(S) comprises summing over all energy contributions of all pairwise atom distances of the system. This implies that many pairs of atom types contribute to the system energy that do not interact significantly with each other. To obtain potentials of mean force $\Delta E_{ab}(d)$ of specific interactions of atom types a and b, the average energy $E(d)$ of all potentials of all atom type combinations at a given distance d is used for calibrating the specific energy potential of a and b. Let n be the number of atom types.

$$E(d) = \frac{2}{n(n-1)} \sum_{i=1,\ldots,n-1} \sum_{j=i+1,\ldots,n} E_{ij}(d) \tag{2.9}$$

$$\Delta E_{ab}(d) = E_{ab}(d) - E(d) \tag{2.10}$$

Insertion of Equation 2.7 and its analog for $E(d)$ in Equation 2.10 yields

$$\begin{aligned}\Delta E_{ab}(d) &= -kT\left[\ln P_{ab}(d)\right] - kT\left[\ln Z_{ab}\right] + kT\left[\ln P(d)\right] + kT\left[\ln Z\right] &(2.11)\\ &= -kT\ln\frac{P_{ab}(d)}{P(d)} - kT\ln\frac{Z_{ab}}{Z} &(2.12)\end{aligned}$$

$P(d)$ is the probability of observing two atoms of arbitrary atom types at distance d. For simplicity, Sippl assumes that the partition functions for all atom pairs are approximatively $Z_{ab} \approx Z$, which causes the second term in Equation 2.12 to disappear. Considering the successes achieved using potentials of mean forces that were reported in recent years, one can conclude that neither this nor the other assumptions made stand in the way of the success of this method. The specific energy of an interacting atom pair a and b can now be formulated as:

$$\Delta E_{ab}(d) = -kT\ln\frac{P_{ab}(d)}{P(d)} \qquad (2.13)$$

If both distance probabilities are given over a continuous sequence of distance intervals this equation defines a potential of mean force for the atoms a and b. The distance probability P_{ab} can be estimated with frequency counts of atoms a and b in certain distance intervals in known near-native conformations of molecular systems. For efficiency reasons, a cutoff for maximum distances is also often defined beyond which no further energy contributions are calculated. By using also the transferability assumption such potentials derived on a representative set of structures can be used to estimate the energy of other molecular systems. The accuracy of a scoring function based on PMFs in a molecular modeling scenario depends on (i) the ability of the structure data set to represent all molecular systems of question, (ii) the chosen separation of energy contributions with a suitable definition of atom types and (iii) the granularity of the potentials, which is given by the bin size x.

Many approaches, like the one proposed by Sippl, implement the calibration of specific interaction energies by averaging over all energy potentials over all pairs of atom types. However, this technique is also criticized, e.g. by Thomas and Dill [207], who argue that such a simplification leads to false and complex distance dependencies between different potentials.

2.2.2 Existing Scoring Functions

This section comprises a description of selected scoring functions that are important for modeling and docking in general but also were involved in the design and evaluation of ROTA. A comparative evaluation of these scoring functions and ROTA is shown in Chapter 3. A comprehensive review on scoring functions was recently given by Rarey et al. [169].

DrugScore

DrugScore is a knowledge-based scoring function for ranking ligand poses in protein-ligand docking. A first version, DrugScore$_{\text{PDB}}$ [63], was derived based on structural data of 6,026 protein-ligand complexes of the PDB and gathered with the ReLiBase system [75, 76] in 2000.

A second version, DrugScore$_{\text{CSD}}$ [214], was published five years later with similar parameters and derivation procedure, but this version was derived on the small molecule crystal packing data (28,642 instances) of the Cambridge Structural Database [3]. DrugScore defines 17 atom types which are quite similar to the atom types of the Sybyl atom-type notation [203] and combine atomic information of element and orbital hybridization. Potentials of mean force are defined for all pairs of atom types and pairwise distances between 1.0 Å and 6.0 Å and single-body potentials of solvent accessible surface (SAS). DrugScore is one of the most important scoring functions to compare ROTA to, since it is one of the first two scoring functions that are based on the PMF formalism and that are applied to the protein-ligand docking problem [199] (the other scoring function is PMF score [151]). DrugScore was evaluated in many comparative evaluations and was successfully applied during virtual screening [48, 67, 116].

ITScore

ITScore [86, 87] is quite similar to DrugScore (based on PMF formalism, application to protein-ligand docking) but uses an iterative derivation scheme similar to that used previously for deriving the ENERGI score [206]: based on 786 structures of protein-ligand complexes from the PDB an initial PMF similar to that of DrugScore (but without a term for energetic contributions of solvent accessible surfaces) is constructed. This PMF is then used to guide the sampling routine of the docking program DOCK [49], which generates alternative protein-ligand complexes. These decoy complexes are then used to derive a new version of the PMF, which is then combined with the old version. This procedure is iterated until the generated complexes are quite similar to the known native structures. The whole procedure therefore has the effect of training the scoring function to support the special sampling routine of DOCK, especially to filter out false ligand poses.

RAPDF

The RAPDF[2] (residue-specific all-atom probability discriminatory function) was developed for discriminating between near-native protein conformations and large sets of decoy protein conformations [184]. As the name indicates, the atom types of heavy protein atoms are defined as amino acid type of the respective atom plus its role in the topology of the respective amino acid. RAPDF was derived based on the conditional probability formalism, which generates potentials that are equal to that generated by the PMF formalism in all respects, except that they are yielding ratios of log probabilities instead of estimates of the free energy. When deriving different version of the RAPDF, the authors showed that a large set of atom types leads to a higher performance of the RAPDF than a simpler and smaller atom type scheme, which also influenced the decision for the ROTA atom type scheme shown later. The RAPDF was successfully applied to quality assessment of generated protein structures on its own but reached even higher discriminatory performance when combined with potentials for backbone torsion angles, atom buriedness and hydrogen bonds [209]. However,

[2] downloaded from the Decoys 'R' Us website at http://dd.compbio.washington.edu/ [183]

when applied to side-chain prediction, it turned out to be inferior to other scoring functions [185].

LUDI and F-Score

The scoring function of FlexX [172], abbreviated *F-Score*, is based on the empirical scoring function LUDI [12, 13] developed by Böhm. LUDI sums over different terms that estimate the contributions of different electrostatic interactions (ΔG_{hb} for hydrogen bonds and $\Delta G_{\mathrm{ionic}}$ for ionic interactions) and entropic (ΔG_{rot}) and hydrophobic effects (ΔG_{lipo}). f is a linear function that penalized deviations from optimum interaction geometries, ΔR denotes distance deviations and $\Delta\alpha$ denotes angular deviations. NROT is the number of rotatable bonds of molecule to be docked:

$$\begin{aligned}\Delta G_{binding} = {} & \Delta G_0 + \\ & \Delta G_{\mathrm{hb}} \sum_{\text{h-bonds}} f(\Delta R, \Delta\alpha) + \\ & \Delta G_{\mathrm{ionic}} \sum_{\text{ionic int.}} f(\Delta R, \Delta\alpha) + \\ & \Delta G_{\mathrm{lipo}} \sum_{\text{lipo}} |A| + \\ & \Delta G_{\mathrm{rot}} \mathrm{NROT}\end{aligned} \tag{2.14}$$

LUDI was fit to the structures of 45 protein-ligand complexes (more structures were later used for training F-Score) and first used in the de-novo ligand design program LUDI. Both scoring functions are especially for scoring protein-ligand complexes that feature many hydrogen bonds, but are less suitable for scoring complexes which are dominated by large hydrophobic interfaces.

2.3 Homology Modeling of Protein Structures

Homology modeling (also known as *comparative modeling*) of proteins creates protein models of certain target proteins by reverting to structural data of proteins that are (ideally) closely related to the target protein. A comprehensive overview on homology modeling was recently given by Dunbrack [42]. Earlier reviews are by Hillisch et al. [77] and Jackobson and Sali [91]. Homology modeling relies on the common observation that the more closely related two proteins are, the more similar are their structures and, to a smaller degree, their sequences. Depending on the availability of one or multiple related proteins and associated structural data as template protein structures, homology modeling can sometimes provide protein models whose quality can match the quality of models that are created using X-ray crystallography. In general, the backbone conformations of related proteins are more similar to each other than this is the case for side-chain conformations [38]. It is therefore a common technique to concentrate on the correct prediction of the protein backbone in a preliminary model and to predict or optimize the conformations of side chains in a second step.

2.3.1 Generation of Complete Protein Models

This section presents the homology modeling programs MODELLER [181] and Prime [193], as these methods are widely used in protein modeling and structure-based drug design and can be potentially extended by methods presented in this work. There exists also a number of Internet-based modeling servers like I-TASSER[3] [232, 241, 242], SWISS-MODEL[4] [158, 191] or Robetta[5] [102] that allow for submitting arbitrary protein sequences to automated or semi-automated protein modeling pipelines. Meta servers like Pcons.net[6] [217] or Genesilico Metaserver[7] [120] allow for distributing a modeling task over a set of such modeling servers and rank the returning protein models. The performance of modeling methods is regularly benchmarked during the Critical Assessment of Methods of Protein Structure Prediction (CASP)[8] [149, 148].

MODELLER

MODELLER[9] generates protein models by iteratively modifying a set of starting models so that they satisfy certain *spatial restraints* [181]. Starting models are assembled from parts of a template structure or multiple superimposed template structures based on a pairwise or multiple sequence to structure alignment with alternative connective chain elements. Spatial restraints are provided first by the relative positions of protein fragments in the alignment and second by basic structural features of proteins, e.g. C_α-C_α distances or main-chain and side-chain dihedral angles, that are represented as empirically determined potentials [182]. Multiple conjugate gradient optimizations [163] are carried out by iteratively applying small rotations about main-chain or side-chain dihedral angles that generate an ensemble of target models that satisfy as many restraints as possible. To avoid getting caught in local minima and for efficiency reasons the optimization procedure starts with optimizing local restraints and includes more and more non-local restraints as the optimization continues. The resulting structure ensembles usually represent the conformational space of the protein in those regions modeled without using template coordinates, whereas protein fragments taken from template proteins mainly reproduce the known conformations of main chain fragments and side chains. A comparative evaluation of ten different modeling programs and techniques showed that MODELLER achieves average backbone quality (about 1.0 Å C_α RMSD for models based on templates with 90% sequence identity to the target protein), whereas the accuracy of predicted side chains is below average [216] (below 50% of modeled side chains have both correct χ_1 and χ_2 dihedral angles if modeled on a template backbone with 90% sequence identity to the target protein). Protein models generated with MODELLER can usually be improved by redirecting side-chain conformations with other programs as it was shown in the same study. Recent versions of MODELLER include different alignment algorithms,

[3] http://zhang.bioinformatics.ku.edu/I-TASSER/

[4] http://swissmodel.expasy.org/

[5] http://robetta.bakerlab.org/

[6] http://pcons.net/index.php

[7] https://genesilico.pl/meta2/

[8] http://predictioncenter.org/

[9] http://www.salilab.org/modeller/

refinement techniques and an interface to Python[10], which renders it as the most extensive public-available toolbox for creation and optimization of protein models.

MODELLER was used for generating multiple homology models of target proteins for flexible docking at the start of this project. However, the observed low accuracy of MODELLER in predicting side-chain conformations impeded docking attempts and so motivated the development of a method for side-chain prediction that can also take side-chain flexibility into account.

Prime

Prime [193] is a protein-structure prediction program (Schrödinger, Inc) that is mainly based on the algorithms of the loop prediction program PLOP [94] and the side-chain prediction program SCAP [92, 93, 233]) (which is described in Section 2.3.2 below). Starting with the (putatively) conserved conformations of α-helices and β-sheets of a template protein, the missing loops are constructed from both ends (except at the chain ends) and joined. The conformational space of the loops is explored exhaustively with a main-chain rotamer library comprising short backbone fragments. Loop conformations fulfilling certain spatial constraints are selected and clustered. Side-chain conformations are then predicted using a slightly modified version of the SCAP algorithm. The resulting all-atom models of the loops are then scored with OPLS force field [98, 100] and the Surface Generalized Born model of solvation [58, 61]. Although these techniques facilitate using Prime for general protein modeling tasks, their main application is modeling the conformational changes of the protein upon ligand binding in the Induced Fit Docking (IFD) procedure of Sherman et al. [193] (see Section 2.4.3).

2.3.2 Side-Chain Prediction

Although all methods presented above provide full protein models with coordinates for all heavy side-chain atoms, there exist a number of reasons that justify the isolated prediction of side-chain conformations using a rigid backbone conformation:

- The general trend of side chains being more flexible than the backbone renders backbone prediction unnecessary if there is already a good model of the protein backbone available, e.g. through a closely related protein in mutation-based protein studies.

- The conformational space of side chains can be represented by rotamers (see Section 2.1.1). This allows for predefining a (comparably) small set of possible conformations before the optimization. This low number of conformations enables pre-computing all rotamer pair interactions, which in turn allows for the early elimination of highly unfavorable rotamers through the Dead-End Elimination (DEE) theorem [36, 64, 65, 122, 133].

- Side-chain prediction is already a hard problem, even if the backbone is held fixed:

[10] see http://www.python.org/ for the Python language homepage

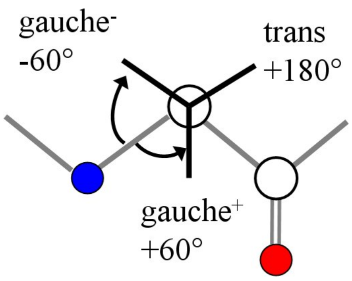


Figure 2.4: Common rotamer states

> the NP-complete problem *satisfiability* (SAT) can be reduced to the side-chain prediction problem (assuming rotameric side-chain conformations and respecting interactions among side chains) [160]. Since the energy score of a protein conformation can be evaluated in polynomial time, the side-chain prediction problem is also NP-complete. One consequence of this is that side-chain prediction can be formulated as a classical combinatorial optimization problem that allows for applying standard algorithms like linear and integer programming [4, 45, 103] or branch and bound techniques [6, 125].

However, through some simplifications (using small numbers of rotamers per side chain, exclusive modeling of short-range side-chain interactions) it becomes computational feasible to calculate the global minimum energy conformation (GMEC), considering only those discrete conformations that can be built with the rigid backbone and a fixed set of rotamers per side chain. Today's side-chain prediction programs can predict the χ_1 dihedral angle of side chains in about 85% correctly, considering two dihedral angles as equal if they are within 40° and modeling on the native backbone. Considering an active site with about twenty relevant side chains that should be modeled for later docking, this results in about three side chains on average pointing in the wrong direction. One possible effect of this is that polar groups that should be oriented towards the ligand are rotated towards the protein, making it impossible for the ligand to establish hydrogen bonds to these side chains. In addition, other side chains that contribute to the protein surface can point into the active site and block the binding of a ligand. Both such effects can render a docking experiment infeasible, which motivates further improvement of current methods.

Rotamer Libraries

This section describes the concepts of rotamer libraries on which the majority of side-chain prediction programs rely. Rotamer libraries are commonly used in side-chain prediction programs to collect all relevant conformations of protein side chains [41], usually for the twenty standard amino acids. Rotamer libraries are derived from protein structure data like that deposited in the PDB and that are preferably of high quality. A rotamer library tries to represent the whole observed population of rotamers (or a major part of them) of the underlying protein data set with a smaller set of rotamers. A common clustering technique is to divide the torsional space of the dihedral angles in three regions, also called slots: *gauge*$^+$ ($\chi \sim +60°$), *gauge*$^-$ ($\chi \sim -60°$) and *trans* ($\chi \sim 180°$) as shown in Figure 2.4.

This coarse cluster scheme groups side-chain conformations that have similar dihedral angles (with angular differences within ±60°). There exist also finer clustering schemes that group side-chain conformations together if their dihedral angles are same within ±40°, 20° and 10° [194, 233].

Usually, the relative frequency of the representative rotamers is also given. The conformation of side chains highly depends on the local backbone conformation and the secondary structure of the local backbone. The SPINFAST approach [161] demonstrates that side-chain conformations can be predicted efficiently using only the target sequence, the template backbone, the secondary structure of the template backbone and the conformations of the corresponding side-chain inside the template protein. Other rotamer libraries contain multiple sets of rotamer representatives that are derived from different structural regions and environments of proteins, e.g. side chains of residues that belong to certain secondary structure elements or side chains of residues which populate a certain region of the Ramachandran Map [167]. One prominent example is the Backbone-Dependent Rotamer Library (BBDep)[11] created by Dunbrack [43]. This rotamer library contains a small set of rotamers for each amino acid (e.g. 3 for serine and 81 for arginine) but has individual rotamer sets for each 10° × 10° sector of the Ramachandran Map. Although the dihedral angles of side chains mostly stay stable between sets of different sectors, the relative frequencies often change drastically. For instance, the relative frequency of the serine side chains having a χ_1 dihedral angle of 66.6° increases from 3% to 87% between the nearby sectors $\Psi =]-160, -150] \times \Phi = [170, 180[$ and $\Psi =]-160, -150] \times \Phi = [120, 130[$. The relative frequencies of this rotamer library are often used as an additional term to empirical scoring functions to model the interaction between side chains and the local backbone segment, for example in SCWRL [26], Rosetta for protein-protein docking [66], the scoring function of Liang and Grishin for side chain prediction [129] and also in this work (see Section 4.6).

The main limitation of a rotamer library is that it just contains the most probable conformations of side chains. Experimental data show that a substantial number of side chains in a protein have an 'non-rotameric' state as their conformations differ significantly from all rotamers in common libraries [190]. This fact puts a natural limitation on the accuracy of all modeling programs that use a certain rotamer library for sampling of the side-chain conformations, which is also the case for the program IRECS [72], that is described in Chapter 4). A side-chain conformation modeled with a rotamer library should therefore be taken as a first guess and should be analyzed for small steric clashes. Eventually, it is meaningful to apply further refinement techniques to such conformations like energy minimization in continuous coordinate space [179, 218] or a combination of different refinement techniques [37].

SCWRL

SCWRL[12] [21, 26] is a side-chain prediction program that guarantees to find an optimal solution – with respect to a simple scoring function – to the side-chain prediction problem and usually can do this within a few seconds (only one out of 160 proteins took over 16 hours, see

[11] http://dunbrack.fccc.edu/bbdep/

[12] http://dunbrack.fccc.edu/SCWRL3.php

Table 4.1). SCWRL uses the BBDep for assigning rotamers to the protein residues. At first, the Goldstein version of the DEE algorithm [64] is used to remove rotamers incompatible with the global minimum energy conformation (GMEC). The interactions of the remaining rotamers are projected onto a graph, on which a biconnectivity analysis is used to divide the graph into components that are then targets for further separate optimization runs. Rotamers are scored by their interaction with the backbone, using a probabilistic score based on the rotamer probabilities of the BBDep, and their interactions with other rotamers, which are scored by a short-ranged steric clash potential (distance cutoff: 3.4 Å). SCWRL (version 3) is popular due to its usability, accessibility, speed and accuracy, and it is probably the most frequently applied stand-alone program for rigid side-chain prediction of proteins in computational biology.

SCAP

SCAP[13] [233] utilizes a set of medium to highly detailed rotamer libraries to sample the conformational space of side chains of a target protein. Side chains are predicted sequentially from the N-terminus of the protein sequence to its C-terminus with repeated runs until no further changes in conformation are reported during a single prediction run. Those rotamers are selected which achieve the best CHARMM score by interacting with the fixed side chains and backbone of the remaining protein. SCAP achieves a similar accuracy as SCWRL does, but requires much more time for optimization.

Side-Chain Prediction with the Self-Consistent Mean Field Approach

The Self-Consistent Mean Field (SCMF) is a fast, heuristic and iterative optimization technique and was first applied by Koehl and Delarue to the side-chain prediction problem [110, 111]. First, all rotamers are assigned to all side chains of a protein on a fixed backbone, using the rotamer library of Tuffery et al. [211]. Then all rotamers are assigned a probability that is used as a measure for preference of each side chain to adopt this rotamer. This probability is distributed uniformly among all rotamers of each side chain before the start of the optimization. Van der Waals (VdW) interactions among rotamers and between rotamers and the protein backbone are scored using a truncated (max. 10 kcal/mol) 12-6 Lennard-Jones potential [127]. The energy of the system is approximated by an effective energy approach, that multiplies each contribution of an interaction by the probabilities of the rotameric states of the interaction partners. During the SCMF optimization the rotamer probabilities are coupled to the effective energy by a modified Boltzmann equation (see Equation 2.3), that determines rotamer probabilities using the contributions of individual rotamers to the effective energy [112]. Both probabilities and effective energies of rotamers are updated self-consistently until the rotamer probabilities converge (typically after about 20 steps). This procedure usually raises a probability of a single rotamer per side chain close to one and reduces the probabilities of the remaining rotamers close to zero, thereby defining a rotamer selection for all side chains. The SCMF approach could achieve an accuracy of 72% for χ_1 and 62% for $\chi_{1,2}$ (both within 40°) on a test set of 30 high-quality

[13] http://wiki.c2b2.columbia.edu/honiglab_public/index.php/Software:Scap

X-ray structures of proteins, which is below the accuracy reported for tools like SCWRL and SCAP.

The SCMF algorithm has been used in place of the Bron-Kerbosch algorithm [23] in FlexE for finding optimal matching sets of side-chain or backbone fragment conformations during ligand docking [71]. The effective energy approach is also used in this work as part of the IRECS algorithm (see Section 4.2.3).

2.4 Protein-Ligand Docking

Protein-Ligand docking is one of the core techniques for structure-based drug design. A docking program can try to dock one or more ligands (within minutes to seconds) into a protein, given that an accurate conformation of the active site is given. Docking ligands usually involves two main tasks. First, the conformational space of the ligand in the active site is explored. This is also the most expensive step in docking as this space is usually extremely large due to the high flexibility of ligands. Second, the binding free energy of each generated complex is estimated by a scoring function. These two tasks are often interwoven since scoring can guide the conformational search of the ligand, and thus reduce the required runtime of the docking program drastically. The high number of putative complexes that must be scored in turn requires that the scoring function can be evaluated quickly. The most important application of docking is the screening of large libraries of chemical compounds for potential drug candidates. This task requires that docking programs are fast and able to test a single ligand within a few seconds.

The following section presents different approaches that either assume that the conformation of the protein does not change upon binding or take protein flexibility into account during docking. Comprehensive reviews on docking were given by Halperin et al. [70], Mohan et al. [146] and Rarey et al. [169].

2.4.1 Docking Approaches Ignoring Protein Flexibility

Current docking approaches ignore the flexibility of the protein if not explicitly stated, whereas ligand flexibility is usually taken into account. This assumption is a strong simplification, since it greatly reduces the conformational space of the protein-ligand complex that must be searched for an optimal solution candidate. The runtime saved can be spent on more sophisticated scoring, taking explicit water molecules during docking into account or for screening larger compound libraries.

The rigid-protein assumption is acceptable if a high-quality protein model is available and it is known that the conformation of this protein does not undergo major changes upon ligand binding. However, even when the protein can change its conformation upon ligand binding, a docking attempt can succeed if the protein conformation is already adapted to the ligand in question or adapted to a whole class of ligands with similar topology and binding mode. Thus, ligands that have different binding modes and topologies compared to the group of known binders – these are of highest interest in pharmaceutical industry – require the adaptation of the protein to these modes for complex formation. These ligands

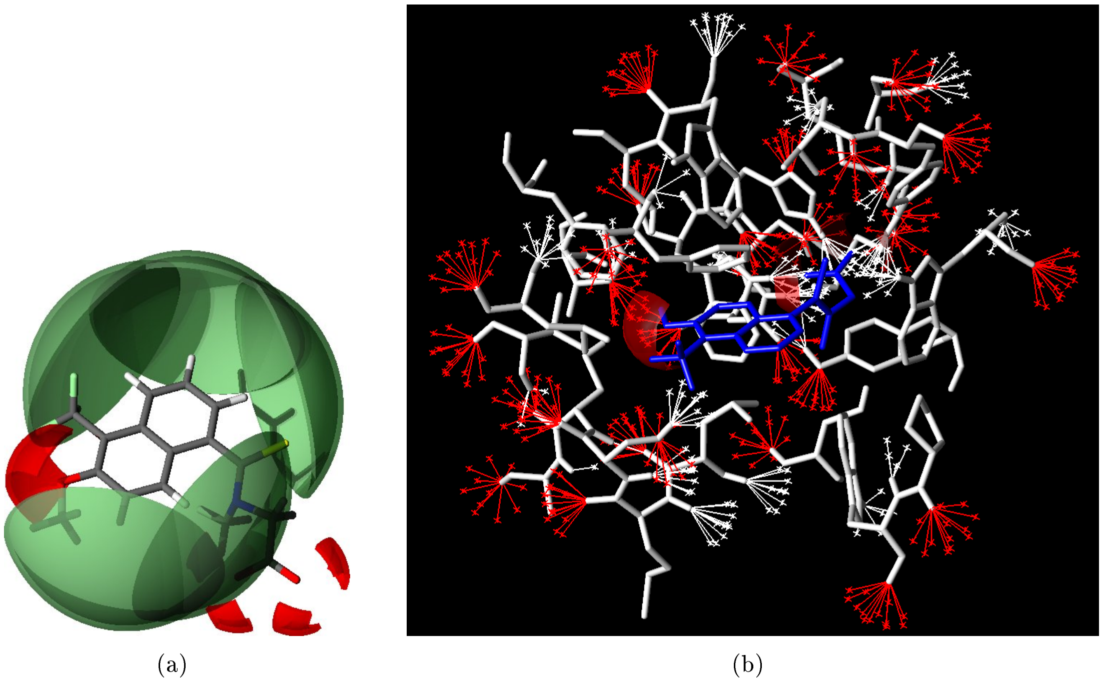


Figure 2.5: Internal representation of protein-ligand interactions in FlexX and FlexE for aldose reductase and the inhibitor tolrestat. (a) The ligand (tolrestat) with interaction geometries for phenyl rings (green), hydrogen bond donors (white) and acceptors (red), extracted from the active site of aldose reductase (PDB ID: 1ah3). (b) The active site (white atoms) of aldose reductase together with tolrestat (blue atoms). Interaction points (protein only) and geometries (ligand only) are drawn for hydrogen bond donors (red) and acceptors (white).

are therefore likely to be docked in a wrong conformation, preventing the prediction of the correct binding mode and affinity.

FlexX

The docking program FlexX[14] [172] and its extension FlexE [33] are used for performing virtual screening experiments in this work, and therefore their algorithms are summarized here (and in Section 2.4.2). FlexX generally follows the incremental construction paradigm: it first cuts the ligand into small fragments. Then, it places a so-called base fragment in the active site of the protein, which is extended by subsequent placements of ligand fragments until the ligand is completely built up in the active site. A modified version of the scoring function LUDI is used for scoring the protein-ligand interactions (see Section 2.2.2).

Many techniques are applied to increase the speed and accuracy of this strategy. After the initial definition of the active site (manual or by a reference ligand position), all atoms of the protein active site are assigned interaction geometries as defined for the LUDI scoring function. Each interaction geometry is approximated by a cloud of interaction points (see

[14] http://www.biosolveit.de/FlexX/

Figure 2.5(b)). Triangles of protein interaction points are stored as line segments in a hash table and are used in the later placement step. These preparation steps are computationally expensive (usually 1-10 minutes), but need to be executed only once per protein and thus only marginally influence the computational cost of a screening experiment (that is: one protein vs. a large number of compounds). The following preparation steps for ligands must be repeated for every ligand that should be docked and therefore contribute primarily to the relevant runtime costs of FlexX during virtual screening. After a query ligand is loaded it is fragmented at rotatable bonds. Interaction geometries are assigned to ligand atoms as previously done for protein atoms. Conformation of molecular rings are computed with CORINA[15] [59, 180].

The first step of complex construction is to place a base fragment into the active site from which the whole ligand is then built up iteratively fragment by fragment according to the order defined through the fragmentation. Base fragments are selected using a number of criteria, e.g. size and number of directional interactions [171]. Triangles are drawn between interaction points of the ligand base fragment and matched to triangles drawn between interaction points of the protein using the hash table. If a match is found a conformation of the respective base fragment is generated. FlexX follows a k-greedy incremental buildup strategy: after step i a set of partial ligand conformations has been generated that are built up to their ith fragment. These partial ligand conformations are then extended in step $i+1$ by adding fragment $i+1$ in various structural alternatives using the MIMUMBA torsion angle database [107]. The resulting set of ligand conformations is then scored using F-Score, and only the k best-scoring placements are retained for subsequent buildup steps. The result is a sorted list of ligand placements with estimated binding affinity to the protein. FlexX is able to generate a native-like ligand conformation (RMSD below 2.0 Å) in about 70% of all proteins in a set of 200 high-quality structures of protein-ligand complexes and can identify such a conformation on top of the sorted list in about 46.5% of these test runs [117].

Other Approaches

Table 2.1 shows the results of a comparative evaluation of the docking programs MVP [121], FlexX [172], Glide [55], Flo+ [140], LigFit [215], Fred [139], DOCK4 [49, 119, 195], GOLD [96, 97], DOCKIt [18] and MOEDock [32] performed by Warren et al. [224]. All programs were used for screening eight different compound libraries with active and inactive compounds for eight different drug targets. FlexX achieved an overall high enrichment factor that was only exceeded by MVP. However, the docking procedure of MVP was built on knowledge of homologous protein-ligand complexes, which was not used for the other docking programs. The results show that there is at least one protein for each docking program at which this docking program performs at best or is among the best programs. This suggests that the choice of the best docking program (and therefore also scoring function) for a virtual screening experiment dependents on the particular target protein. Actually, many virtual screening experiments are performed by running several docking programs.

[15] http://www.molecular-networks.com/software/corina/corina_f.html

Table 2.1: Enrichment factors achieved on eight drug targets by ten docking programs

Docking program	average	Chk1	FXa	GB	HCVP	MRS	E. coli PDF	Strep PDF	PPAR δ
ideal	9.2	10.0	9.8	10.0	9.5	10.0	7.6	8.3	8.6
MVP	5.7	7.2	5.8	5.3	3.6	6.4	6.7	6.9	3.9
FlexX	3.3	7.0	2.2	5.8	0.9	3.9	0.8	0.8	5.2
Glide	2.9	6.3	3.4	1.0	1.0	5.3	0.6	0.4	4.8
Flo+	2.7	5.6	2.7	2.3	3.4	1.7	1.5	0.8	3.6
LigFit	2.3	3.3	1.9	2.8	1.8	2.9	2.9	1.7	1.2
Fred	2.1	2.9	4.1	1.9	2.0	0.6	3.2	1.2	1.1
DOCK4	2.1	1.4	4.1	1.7	1.8	4.2	0.9	0.8	1.7
GOLD	2.0	0.1	4.1	4.0	0.0	0.8	1.0	0.1	5.5
DOCKIt	1.7	4.2	2.0	2.0	1.0	1.0	0.2	0.0	3.2
MOEDock	1.0	3.9	0.6	0.0	0.0	1.0	2.1	0.6	0.0

Comparative evaluation of ten docking programs by virtual screening performed by Warren et al. [224]. Each docking program was used to dock active and inactive compounds. The enrichment factor was computed by measuring the relative enrichment of active compounds at the top 10% of the sorted (by docking scores) compound lists. Abbreviations: Chk1: checkpoint kinase 1; FXa: factor Xa; GB: gyrase B; HCV: hepatitis C virus protease; MRS: methionyl-tRNA synthetase; PDF: polypeptide deformylase; E. coli: escherichia coli; Strep: streptococcus pneumococcus; PPARδ : peroxisome proliferator-activated receptor δ. Adapted from Warren et al. [224].

2.4.2 Docking Approaches Incooperating Protein Flexibility

Among the many challenging aspects of protein-ligand docking, this work concentrates on dealing with protein flexibility and with inaccurate positions of side chains during docking. An increasing number of papers also dealing with these issues (see below) published recently show that these topics are of major interest and that no universal solution has been found yet. Among the approaches that try to incorporate protein flexibility many are extensions to already existing docking programs (e.g. FlexE to FlexX, IFREDA [30] to ICM [1]), but there exist also programs that were designed specifically to account for protein flexibility (FLIPDOCK [243], SLIDE [189, 237]).

This section highlights such approaches that are either used in this work or that are likely to benefit from methods which are developed in this work. More comprehensive overviews on state-of-the-art approaches in this field are also available in the literature [27, 28, 204].

Serial Docking

The most self-evident strategy to account for protein flexibility during docking is to repeat the traditional docking procedure (rigid protein – flexible ligand) with a set of different protein conformations. This technique depends on an external source of protein conformations – these can originate either from multiple X-ray structures, from molecular dynamics simulations or from special sampling techniques. The docking experiments performed by Frimurer et al. into tyrosine phosphatase 1B [56] can illustrate this technique: they started with a crystal complex of the protein and a small peptide. They already knew from structural analysis that only 3 side chains are flexible in the active site and they could cut down the number of relevant rotameric states of them to 3, 4 and 8, respectively. This yielded a number of 96 different protein conformations in which all possible combinations of rotamers are present. They collected three known inhibitors for which a complex structure with tyrosine phosphatase was available and docked them into each of the 96 protein models and the protein conformation of the peptide complex. They found that the crystallized ligand conformation could be predicted at best by docking them sequentially into the 96 protein models and preserving the lowest energy conformation of each ligand.

FlexE

FlexE uses ensembles of superimposed X-ray structures of a single protein to capture the boundaries of the conformational space of the protein. FlexE cuts these structures at the peptide bonds and the connecting bonds of side chains and backbone into segments. All conformational variants – called instances – of the protein segments are then combined in a so-called *united protein description* [33]. This data structure is prepared for docking just as proteins are prepared in FlexX (see Section 2.4.1). Ligand fragments are also placed and extended like in FlexX with the exception that for each generated ligand conformation an adapted conformation of the protein is generated by selecting from the segment instances with the SCMF algorithm (see Section 2.3.2) [71]. FlexE is able to improve the docking accuracy on protein targets that undergo conformational changes upon binding by the cost of extended runtime.

2.4.3 Docking into Homology Models

Inaccurate positions of protein atoms are among the most frequent reasons for a docking experiment to fail. Such inaccuracy causes deformations of the internal representation of the protein surface and distortions of interaction geometries that can spoil the docking process [39, 50, 99, 153]. Since such misplaced atoms are also frequent in X-ray structures, docking tools are usually able to cope with a limited amount of atom position inaccuracy. FlexX for example has many parameters that allow for manually adjusting the tolerance cutoffs for overlaps between ligand and protein and also angular tolerances for building hydrogen bonds. Setting tolerance parameters too generously however also incurs the risk of creating irrelvant complex conformations. If this happens during a screening experiment, many false protein-ligand complexes are created and it is much harder for the scoring function to identify the true positive binders from the mass of false positive binders.

One example is the series of experiments performed by Thorsteinsdottir and coworkers in which they compared docking into X-ray structures and inaccurate models of HIV protease [208]. They calculated binding free energies from molecular dynamics simulations, using first X-ray structures of the complexes and second protein models built from the backbone of these X-ray structures and side-chain conformations built with SCWRL. They found that the distortion which resulted from the inaccurate remodeling with SCWRL decreased the correlation coefficient between experimental and calculated binding free energy from 0.9 to 0.75.

Soft Docking

Soft docking is an approach that enables a docking program to tolerate small geometric errors by reducing the usually high penalties for such errors and therefore 'softening' the potential walls that surround the minimum energy conformation of the protein-ligand complex [51]. Bindewald et al. used an automated optimization procedure for developing a scoring functions based on force fields that is composed of weighted terms for van der Waals interactions, hydrogen bonds, electrostatic interactions and internal ligand energy [17]. They show that the parameters of the function optimized for docking into structures of low resolution produce much smoother potentials with much lower potential walls at close atom distance than those parameters that are optimized for docking into structures of high resolution.

MOBILE

The MOBILE [46] approach of Evers et al. interweaves the homology modeling process with the docking of known ligands. At first, a rough model of the target protein is generated with MODELLER by using only the structure of the template protein(s). A docking of all known ligands (preferably ligands of the target protein, but ligands of the template protein can also be used sometimes) into these protein models is carried out with the docking program AutoDock [154, 147]. These preliminary models of protein-ligand complexes are then further optimized with MODELLER. Modified DrugScore potentials are incorporated in the force field of MODELLER to enable MODELLER to account for the interactions of the protein with the (fixed) ligand conformation during the generation of optimized protein models. The DrugScore scoring function is then used to rank the resulting model and guide the final per-residue optimization that leads to a single resulting model. The approach was successfully tested on modeling and docking to G-protein-coupled receptors (GCPRs), which represent an extraordinarily hard task for both homology modeling and docking [47, 48]. A similar approach developed in this group, DRAGHOME has been successfully tested on Thrombin [187].

IFD

The Induced Fit Docking (IFD) procedure of Sherman et al. uses the protein-structure prediction program Prime [193] (see Chapter 2.3.1) and the docking program Glide [55, 69] (both Schrödinger, Inc.) to fit the active site of a protein to a given ligand. In a first step the ligand is docked with Glide into the unadapted active site. To minimize the risk of steric clashes with the unadapted side chains during docking the potentials of the scoring functions are made 'soft', that is, the penalty score for short-range electrostatic repulsion is lowered. For the same reason the side chains of the initial model are checked for flexibility by comparison with respective residues in superimposed X-ray structures of the same protein in holo form (bound to different ligands). If a side chain is flexible, its atom representations are temporarily deleted, except the C_β atom, to minimize the chance of clashes between the ligand and the flexible part of the side chain. Twenty different alternatives of the resulting complexes are then further optimized by Prime, now using also the full conformations of flexible side chains as predicted with SCAP algorithm [92, 93, 233]). The ligand is then docked a second time with the standard Glide algorithm (no softening of potentials) into the active sites of those optimized protein models that achieved a Prime score below a certain threshold. The resulting complexes are then assigned a final consensus score made up from the Glide and Prime scoring functions. Finally, the best performing complex is returned as the final structure or – if no complex achieves a score below a second threshold – the whole procedure is executed a second time with less softened potentials in the initial docking step. Compared to MOBILE, the IFD approach expands multiple placement options for the ligand in its optimization procedure, which renders it more tolerant to errors in the initial ligand placement and complex scoring. IFD showed promising results in a redocking experiment on a small data set [193].

Chapter 3

The ROTA Scoring Function: Soft Potentials of Mean Force for Intra-Protein and Protein-Ligand Interactions

This chapter describes the derivation procedure of the ROTA scoring function and presents the results of two comparative evaluations. The derivation procedure of ROTA follows the Potential of Mean Force (PMF) approach described in the introduction (see Section 2.2.1) that requires the following steps: (i) creation of structure libraries, (ii) calculation of distance distributions, (iii) generation of discrete probability distance functions and (iv) derivation of discriminatory potentials of mean force. The ROTA scoring function was derived for guiding the side-chain prediction of IRECS and also the protein-ligand docking of FlexE. One structure library *NLIB* that is used for deriving ROTA therefore comprises native structures of proteins and protein-ligand complexes. A second structure library *DLIB* contains decoys that are generated using the BBDep rotamer library and FlexX – similar to the derivation procedure of ITScore [86, 87] – to catch the sampling errors that are likely to occur using these methods. In a later evaluation it is shown that ROTA potentials are able to successfully discriminate between erroneous (different from conformations observed in X-ray structures) and near-native conformations of side chains and ligand poses. Although no data about binding affinity were used in the derivation procedure, ROTA can also estimate the binding free energy of a protein-ligand complex with high accuracy by summing over the contributions of all interactions between a ligand and a protein (see Section 2.1.2). The latter application is tested on a set of protein-ligand complexes with experimentally determined binding affinities which allows for comparing ROTA with fourteen other scoring functions.

The following sections introduce the structure libraries that were constructed for deriving ROTA. ROTA can also be derived using the conditional probability formalism introduced by Samudrala and Moult [184] as it was shown previously [72]. This formalism is rarely chosen in the field and does not provide a direct relationship between log probabilities and contributions to the system energy as the PMF formalism does. Therefore in this work it

Table 3.1: Properties of the ROTA decoy sets for side-chain prediction

Test decoy sets	# Proteins	# Decoys	Side-chain RMSD range [Å]
4state	7	4656	2.55 - 13.25
All-BBDep	10	5000	0.76 - 3.68
All-Rot	10	5000	1.81 - 4.34
Triple-BBDep	10	1842	0.01 - 3.06
Triple-Rot	10	1842	0.06 - 3.15

was decided to use the alternative PMF derivation procedure. However, Samudrala and Moult pointed out that both formalisms are equivalent for practical uses.

3.1 Generation of the ROTA Structure Libraries for Side-Chain Prediction

The structure libraries for side-chain prediction are based on a preconfigured set of protein structures, the top 500 database [135], downloaded from the Richardson Lab web page[1]. This set of protein structures was assembled by Lovell et al. [135] for building a high-quality representation of the structural variety of the PDB. The set contains only X-ray structures that have a resolution of 1.8 Å or better, low B-factors, and passing multiple geometry checks (e.g. for clashing atoms or unfavorable backbone torsions). All chains have lower sequence identity than 30%. All hydrogen atoms and atoms of hetero compounds are omitted so that the protein structures of the two structure libraries contain only heavy atoms of standard amino acids. From this set ten structures were randomly (structure probabilities were distributed uniformly) set apart for later testing purposes. This set is named the MQAP test set (PDB IDs: 1edg, 1jet, 1mb4, 1pen, 1qb7, 1qq4, 1vfy, 1wab, 3ebx, 3ezm).

For each of the 500 proteins a number of decoys structures were generated. However, only decoys from the 490 remaining proteins were used for training the other decoys were added to the MQAP test set. Four different decoy sets for the structure library DLIB were created, since we wanted to try different methods of generating decoys for the DLIB: All-BBDep, All-Rot, Triple-BBDep and Triple-Rot. Each decoy set was generated using a different procedure to capture different aspects of fallacious side-chain prediction. For the All-BBDep and All-Rot sets all side chains were randomly rotated simultaneously. This procedure was performed ten times for each protein structure, yielding ten different decoy structures for a single native structure. The decoys of the sets Triple-BBDep and Triple-Rot were created by rotating the side chains of only three neighboring residues simultaneously. This procedure generated decoys that were more similar to the native structures and therefore harder to detect.

The side-chain dihedral angles of the decoys in the All-Rot and Triple-Rot sets are simply

[1] http://kinemage.biochem.duke.edu/

distributed uniformly over the whole range of dihedral angles between −180° to 180°. For the sets All-BBDep and Triple-BBDep rotamers were drawn from the BBDep, with a selection probability among all available rotamers as defined in the library (see Section 2.3.2). This causes the decoys to contain mostly the rotamers favored by the BBDep, but ignores any interactions among rotamer pairs. Since these decoys will be later included in the derivation of ROTA, ROTA favors rotamers that are less probable in the BBDep. This setup was chosen to adjust the ROTA scores to the BBDep frequency scores, as both scores should be later combined for the scoring of rotamers in the side-chain prediction tool IRECS (see Section 4.2.2).

Steric clashes were not removed from the generated conformations, since close distances should also be contained in the ROTA potentials so that ROTA can penalize clashes without the help of other scoring schemes. As a result, most of the decoys are clearly identifiable as decoys by quite simple methods. Only in the case of small perturbations, which are frequent in the Triple-BBDep and Triple-Rot sets, decoys are sometimes hard to distinguish from the starting crystal structures. Table 3.1 lists the number of proteins, decoys and side-chain RMSD ranges between decoys and proteins. The decoy set *4-state* [155] is used during the evaluation and is described in Section 3.6. The results of this evaluation determined the decision to exclusively use the All-BBDep version of the DLIB for further scoring of rotamer-rotamer and rotamer-backbone interactions.

3.2 Generation of the ROTA Structure Libraries for Docking

The structure libraries used for deriving the ROTA potentials for side-chain prediction cannot be used for the derivation of potentials for docking, since these structure libraries contain too few protein-ligand complexes and the ligand positions do not differ between structures in NLIB and DLIB. Therefore, two additional structure libraries had to be derived on a different structural basis.

This structural basis was drawn from the PDB [10] using a filtering and clustering procedure. This procedure is summarized in the upper part of Figure 3.1: at first, a set of relevant structures is extracted from the PDB. Then, these structures are clustered, so that a smaller set of representative structures is retained for all relevant structures (4778 structures in total). Finally, all structures of the representative library are modified in two ways to generate NLIB and DLIB. ROTA is supposed to empower FlexE to dock into structures with modeled side chains. This requires structure libraries that are representative first (NLIB) for protein-ligand complexes and second (DLIB) for failures in (i) modeling of flexible side chains and (ii) docking into flexible proteins with FlexE. ROTA should therefore be able to tolerate small to medium positional errors of side-chain atoms.

3.2.1 Extraction of Relevant Complexes

Only a fraction of the structures deposited in the PDB are usable for deriving potentials of mean force. The following constraints and cutoffs are commonly used in the field [63, 151, 184] to filter relevant and representative subsets of the PDB and were not subject to any optimization attempts. A protein-ligand complex was only selected from the PDB if

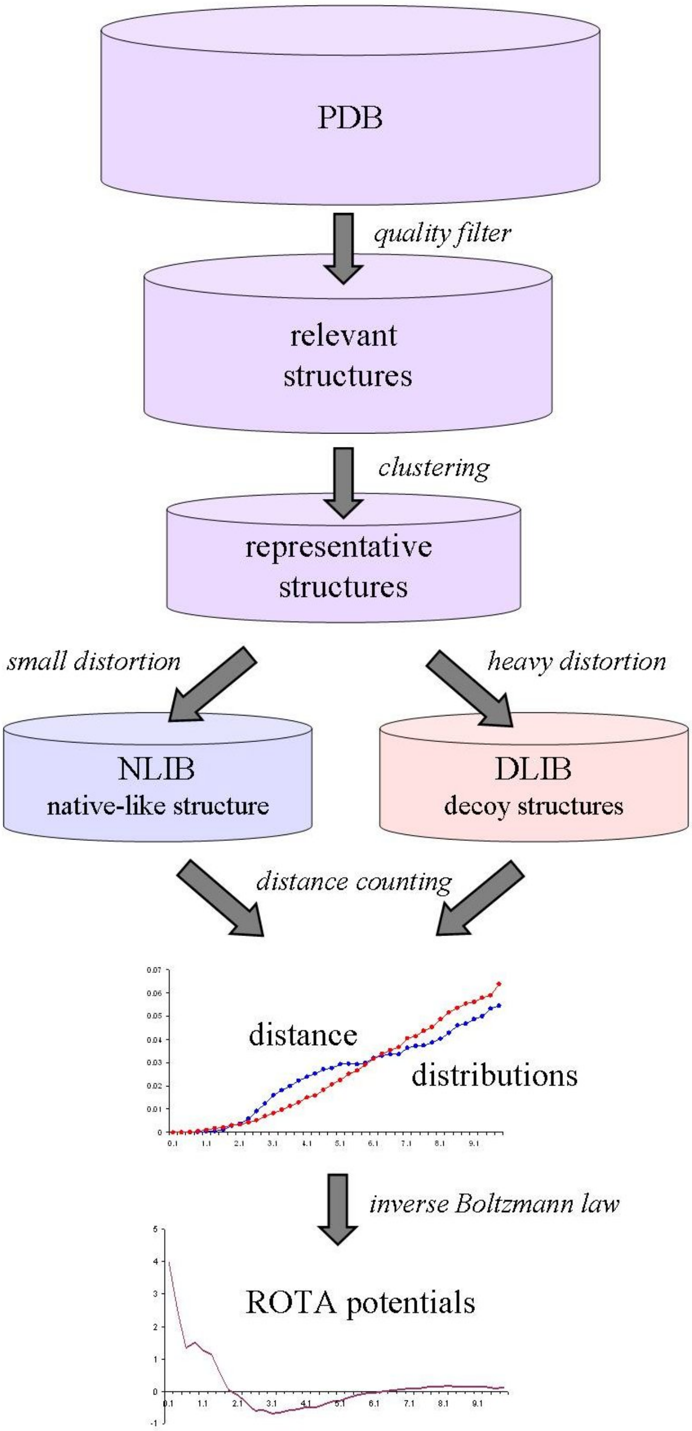


Figure 3.1: ROTA derivation procedure

1. the structure has been resolved by X-ray crystallography with a resolution better than 2.5 Å,

2. the structure contains a ligand which has between 6 and 36 atoms,

3. this ligand is not covalently bound to the protein and

4. the protein consists of more than 20 residues.

The first restriction guarantees a certain quality level of the structural data. The threshold is intentionally set high, since ROTA should also be able to cope with small-scale errors made during modeling of X-ray structures. The second restriction filters out protein-ligand complexes that do not contain a drug-like ligand. The chosen threshold however allows for peptides or peptidomimetic ligands with about five residues. The reason for the third restriction is that distance counts from covalently bonded atoms must not be mixed with distance counts from non-covalently bonded atoms. This is necessary since a mixture of both kinds of interaction would result in nonspecific potentials. The final restriction ensures that only such proteins are selected that have a chance of becoming drugable targets, because proteins require a minimum size for building drugable structures like surface clefts or binding pockets with sufficient stability.

3.2.2 Extraction of Representative Complexes

A prerequisite of applying the Boltzmann law for the derivation of potentials of mean force for distances between atom groups is that the frequencies with which such distances occur within the structure library are representative for the energetic states of conformations of the particular protein-ligand substructures (see Section 2.7). This requirement is not fulfilled by the protein structures deposited in the PDB since the PDB contains different protein families in wildly different densities. The structural features of hemoglobin for example are quite frequent in the PDB just because hemoglobin is such an important protein for medicinal chemistry, and not because these features have an exceptionally low energy state.

Two filtering procedures were applied for removing redundancy of protein structures among the relevant complexes. The first filter is designed to prevent certain frequent ligands from dominating the data set. For this purpose, groups were formed out of all complexes that bind the same ligand (identified by its name as found in the PDB file). Each group was checked if it contains protein pairs with a higher sequence identity than 50% (checks were performed in the alphabetic order of PDB IDs). If this was the case, an arbitrary protein of the two conflicting ones was removed from the group. This procedure was repeated until there were no such protein pairs remaining in the groups. Sequence identity was checked with precompiled clusters that were derived with the program CD-HIT [128] using a sequence identity cutoff of 50%. These clusters are available for different sequence identity cutoffs in the 'derived data' section of the PDB[2]. The resulting data set contains 4,778 X-ray structures of protein-ligand complexes. See the appendix C.5 for a list of the selected PDB structures.

[2] ftp://ftp.rcsb.org/pub/pdb/derived_data/NR

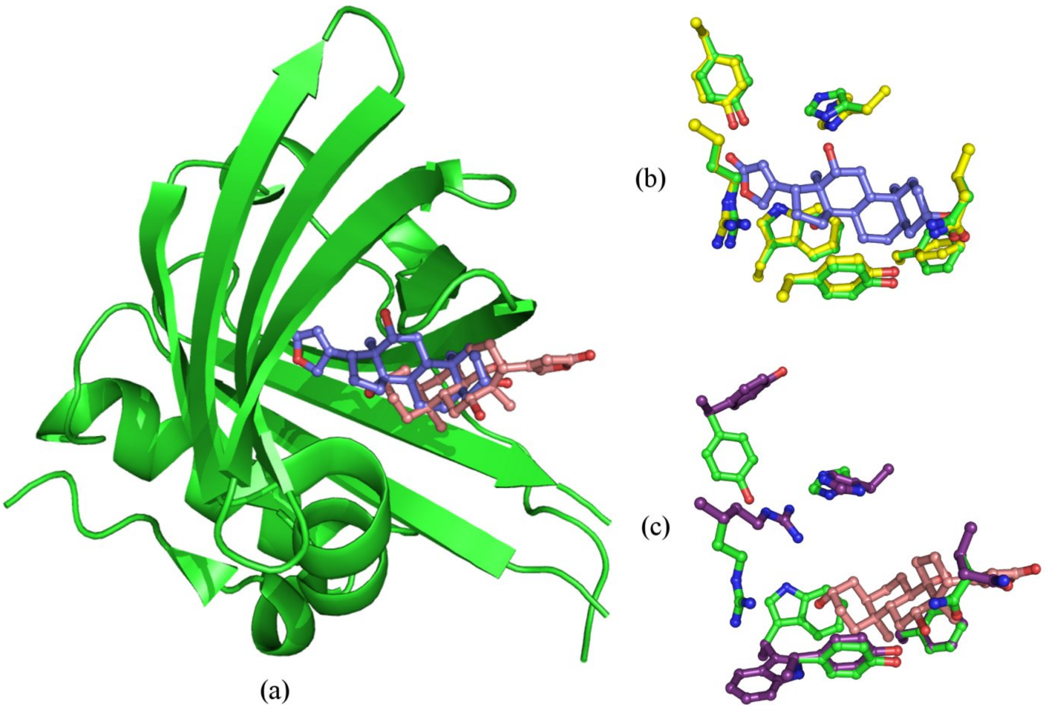


Figure 3.2: Distortion of a complex conformation. On the left hand side (a) the complex structure of the protein lipocalin (green, original x-ray structure [114]) with its ligand digoxigenin (blue) is shown together with the decoy conformation (red) of the ligand. On the right hand side the modeling procedure for structures in the NLIB and DLIB structure libraries are visualized. In (b) the native conformations of the protein side chains (green) are exchanged with the closest available rotamers in the rotamer library (yellow) for the NLIB. In (c) the native conformations of the protein side chains are exchanged with random rotamers from the rotamer library (violet) for the DLIB.

3.2.3 Generation of Complexes for NLIB

The set of structures in NLIB is not just a representative set of X-ray structures, as described for the structure libraries for side-chain prediction (see Section 3.1), but a perturbation technique is applied to the representative crystal structures to soften the later derived potentials of mean force: all side chains of the protein structures are replaced by the best matching (considering dihedral angles) rotamer from the BBDep rotamer library, whereas the ligand conformations and positions are conserved. This modification of X-ray data is supposed to generate models that are most similar to those complexes which will later be scored by ROTA and have their side-chain conformations taken from the BBDep. The ligand positions are perturbated by redocking the ligands into the X-ray structures with FlexX. This tasks is eased by adding the value $x * min(dist(n, c), 4.0)$ to the score of each intermediate docking solution of FlexX, whereas $x = 10$ and $dist(n, c)$ is the RMSD between the native ligand conformation (n) and the respective conformation of the intermediate docking solution (c).

However, the default workflow (including determination of the active site, surface computation, protonation, template assignment and docking algorithms) of FlexX was taken in its generic fashion and was not adapted to the individual proteins. Using this procedure a ligand conformation with an RMSD of 2.0 Å or better could be reproduced in 3,070 cases (of 4,778 complexes). In all other cases the ligand conformation from the original X-ray structure is retained instead of a docking solution.

3.2.4 Generation of Complexes for DLIB

The set of decoy structures is also created from the set of representative protein structures. The ligands are extracted from the binding pockets and redocked with FlexX, but to confound the docking process the FlexX score is modified with the procedure as described above in Section 3.2.3, but this time with x set to -10. This effectively forces FlexX to misplace the ligands on the protein surface or the boundary of the binding sites, which gives a representative set of worst-case results that FlexX can generate. Afterwards, the conformations of protein side chains are randomized with the same procedure described in Section 3.1 for the decoy set All-BBDep. Finally, the protein complex with maximal RMS deviation from the crystal structure concerning the ligand conformation is added to the DLIB structure library. In the 133 (of 3,070) cases where FlexX still manages to dock the ligand with an RMSD below 3 Å, the native conformation of the ligand is used instead of any decoy conformation. These are rare cases and do not require special treatment since the disturbed side-chain conformations of the proteins also assure that no accurate geometrical data can enter the DLIB structure library. Figure 3.2 shows an example protein-ligand complex with the two types of generated complexes for the NLIB (upper right) and DLIB (lower right) structural libraries.

3.3 ROTA Atom Types

As described in Section 2.2.1 one of the most important properties of a knowledge-based scoring function that uses potentials of mean force for atom distances is the applied atom type scheme. One fundamental design decision for ROTA was to define different sets of atom types for proteins and ligands, as it was done previously for the DFIRE scoring function [238, 239, 240]. In an early version of ROTA there were also atom types defined for hydrogen atoms. However, later evaluations of ROTA showed that these additional atom types did not significantly improve prediction accuracy, but drastically increased the computational effort to calculate interaction scores for heavy atoms and hydrogen atoms. Therefore, the scoring of hydrogen atoms was omitted in the current version of ROTA.

3.3.1 Ligand Atom Types

In order to derive potentials of mean force with sufficient quality, there must exists a sufficient number of pairs among all atom types that establish specific interactions (and are therefore close to each other). Thus, the atom types for ligands must be chosen coarsely, so that they

Table 3.2: ROTA atom types for ligand atoms

Atom type	Description
C.AR	aromatic carbon
C.O2	carbon in carboxylate group
C.O1	carbon with one oxygen bound
C.N1	carbon with one nitrogen bound
C.N2	carbon with two nitrogen atoms bound
C.H0	carbon with no hydrogen atoms bound
C.H1	carbon with one hydrogen atom bound
C.H2	carbon with two hydrogen atoms bound
C.HX	carbon with more than two hydrogen atoms bound
N.AR	aromatic nitrogen
N.AM	amide nitrogen
N.H0	nitrogen with no hydrogen atoms bound
N.H1	nitrogen with one hydrogen atom bound
N.HX	nitrogen with more than one hydrogen atom bound
O.CO2	oxygen in a carboxylate group
O.H1	oxygen with one hydrogen atom bound
O.C	oxygen bound to a carbon
O.X	oxygen bound to something else
S	sulfur
HAL	halogen atoms
X	everything else

These atom types are used for grouping ligand atoms by their putative non-covalent interactions with proteins. These atom types are also used as a second, coarser set of atom types for protein atoms.

fragment the diverse space of ligand atoms into large clusters. The mapping of protein-ligand interactions to atom types must therefore remain quite simple. Like many other scoring functions, ROTA is based upon the atom types of the SYBYL Mol2 format[3]. These atom types are mainly a composition of element, charge and hybridization state of atoms. Some additional atom types are introduced to capture frequent substructures in organic chemistry (amide, aromatic and carboxyl groups). The hybridization state of an atom is primarily important for building the 3D geometry of a molecule, whereas the interactions with other atoms are mainly determined by element, partial charge and bound hydrogen atoms. Therefore ROTA omits the hybridization state of an atom, but adds the number of bound hydrogen atoms to the element or substructure information, as it is also done for the atom types of ITScore [86, 87]. The resulting atom types are shown in Table 3.2.

[3] http://www.tripos.com/data/support/mol2.pdf

3.3.2 Protein Atom Types

ROTA defines two different sets of atom types for protein atoms. The names of the atom types of the larger set A (167 types), are the same as defined for the ROTA potentials for side-chains prediction, that is, the three-letter-code of the amino acid type plus atom name, as found in the respective PDB file. This choice was made due to the good results that were reported by Samudrala and Moult [184] for the RAPDF scoring function (see Section 2.2.2). This atom type set captures the fine differences in interaction profiles of protein atoms that are hard to determine using chemical profiling. For example, the C_α atoms of alanine are more frequent in the core of the protein than the C_α atoms of serine due to the different chemical nature of their side chains. Therefore their preferences for nearby atoms differ significantly although these atoms are chemically similar.

The atom types of the second set B are the same as for ligand atom types, see the Section B in the appendix for a mapping of PDB atom names to ROTA atom types. This second atom type set for protein atoms is used for deriving alternative interaction potentials for protein-ligand interactions. This became necessary since there were often too few small atom distances between ligand and protein atoms to enable the derivation of certain interactions potentials between rare protein and ligand atoms. Therefore, the second set of atom types allowed for creating a much smaller number of protein-ligand interactions potentials (21 × 21 for atom type set B compared to 167 × 21 possible potentials for atom type set A), for which more distance counts of protein and ligand atoms were available. If there are multiple potentials defined for certain pairs of protein and ligand atoms, the potentials based on the atom type scheme A are always preferred, since these sets allow for a much more detailed modeling of protein-ligand interactions.

3.4 Derivation of the ROTA Potentials

The final steps of the derivation procedure of the ROTA potentials are shown in the lower part of Figure 3.1. Two distance distributions with bin size x of 0.25 Å and a maximum distance of 10.0 Å are derived from the structure libraries NLIB and DLIB for all pairs of atom types (ligand and protein sets A and B). As known from the literature [63, 86, 143, 151, 184, 196], distance distributions with too few counts are not suitable for derivation of potentials of mean force, and therefore a distance distribution is only used for potential derivation if it contains more than 500 counts. Otherwise, no ROTA potential for a particular pair of atom types is derived. This happens especially for certain combinations of protein atom types from type set A and rare ligand atom types, e.g. ligand halogen atoms and tryptophane atoms. The definition of two atom type sets A and B for protein atoms in the previous section allows to yield meaningful interactions scores for some of these low count cases.

From these distance distributions the relative frequencies F_{ab} are derived for all atom types a and b. This is done by dividing each count number N_{ab} of atom types a and b from a distance interval of size x by the total number of counts N of the respective distribution:

$$F_{ab}\left(\left[d-\frac{x}{2},d+\frac{x}{2}\right[\right) = \frac{1}{N}N_{ab}(\left[d-\frac{x}{2},d+\frac{x}{2}\right[) \qquad (3.1)$$

It is assumed now that 500 counts are sufficient to convert the relative frequencies directly into distance probabilities by applying Equation 2.8. The distance probabilities of the NLIB structure library are denoted with $P^{+}(d)$, the distance probabilities of the DLIB structure library with $P^{-}(d)$ and a mean distance probability, which is averaged over all probabilities of all atom types for a certain distance d, with $P(d)$. With these distance probabilities it becomes possible to derive two different potentials of mean force using Equation 2.13.

$$\Delta E_{ab}^{+}(d) = -kT \ln \frac{P_{ab}^{+}(d)}{P(d)} \tag{3.2}$$

$$\Delta E_{ab}^{-}(d) = -kT \ln \frac{P_{ab}^{-}(d)}{P(d)} \tag{3.3}$$

$$\tag{3.4}$$

The potentials ΔE_{ab}^{+} are able to identify native structures whereas the potentials ΔE_{ab}^{-} are able to identify decoy structures. The ROTA potentials are now derived as discriminatory potentials from these potentials:

$$S_{ab}^{ROTA}(d) = \Delta\Delta E_{ab}(d) = \Delta E_{ab}^{+}(d) - \Delta E_{ab}^{-}(d) = -kT \ln \frac{P_{ab}^{+}(d)}{P_{ab}^{-}(d)} \tag{3.5}$$

The ROTA potentials are therefore only derived from the distance probabilities $P^{+}(d)$ and $P^{-}(d)$, the mean distance probability $P(d)$ cancels out. From now on the ROTA scores are given in units of kT.

Equation 3.5 can only generate ROTA scores for distance bins for which both distance probabilities are larger than zero. However, for small distances there are often no samples for one or both distance distributions, since clashing atoms are rare in both sets (although structures of DLIB generally contain more clashes than structures of NLIB). Other scoring functions based on potentials of mean force deliberately ignore these distances and are only defined for atom distances above certain thresholds (DrugScore: 1.0 Å [63], BLEEP: 2.5 Å [143, 144]). As the availability of counts at close distances depends on the particular pair of atom types, it was decided to assign a constant penalty to all distance bins for which either $P^{+}(d)$ or $P^{-}(d)$ could be computed. For these distances the ROTA potential is just set to the worst (highest) ROTA score among all potentials, which is 4 kT (rounded up).

Atom distances that should be scored with ROTA usually deviate from the midpoints of count bins that were used to derive the distance distributions. The ROTA score for a distance d between the midpoints d_l and d_r of two distance bins is therefore linearly interpolated between the scores $S(d_l)$ and $S(d_r)$ of these distance bins (adapted from Samudrala and Moult [184]):

$$S(d) = S(d_l) + \left[(S(d_r) - S(d_l)) \frac{d - d_l}{d_r - d_l} \right] \tag{3.6}$$

If a (small) distance has no left midpoint, its score is set to the right midpoint, and if a (large) distance has no right midpoint, its score is set to the left midpoint, as long as this distance does not exceed 10 Å.

3.5 Properties of the ROTA Scoring Function

The ROTA potentials have special properties that cannot be found in a similar combination among other scoring functions. The most important feature of ROTA is that the potentials are soft, so they have low penalties for clashing atoms and wide potential wells for favorable interactions. Figure 3.3 shows the PMFs of $\text{DrugScore}_{\text{PDB}}$ and $\text{DrugScore}_{\text{CSD}}$ for N.pl3 and O.co2 Sybyl atom types that can be compared with the ROTA PMFs for the atom type pairs (HET-O.CO2, LYS-NZ) in Figure 3.4(b) since both are atom types for hydrogen-bond donors and acceptors. The largest difference between the PMFs of ROTA and $\text{DrugScore}_{\text{CSD}}$ is that ROTA penalizes clashes in the distance interval of 0.0-1.3 Å, in which $\text{DrugScore}_{\text{CSD}}$ assigns no penalty scores. $\text{DrugScore}_{\text{CSD}}$ strongly penalizes the distance interval of 1.5-2.5 Å and then rapidly changes to beneficial scores for the optimal hydrogen bond distance of about 2.8 Å for these atom types. The same minimum scores can be found in the ROTA potential of (HET-O.CO2, LYS-NZ). However, this potential is much softer, as the transition between clash scores and scores for the optimal bond distances is spread over the whole distance range between 1.3 Å and 2.8 Å. One implication of this difference is that ROTA is better suited for scoring complex conformations in which rotamers of lysine are not adapted to a certain ligand conformation, because there is a high chance that the distance between a ligand hydrogen acceptor and the N_ζ of lysine is smaller (or larger) than the optimal distance. $\text{DrugScore}_{\text{CSD}}$ does not tolerate such a case and will penalize this distance, although the general conditions for a hydrogen bond may be met. The remaining parts between 3.0 and 6.0 Å of the ROTA and DrugScore potentials are quite similar.

The ROTA potential for (HET-O.CO2, LYS-NZ) has a small artifact at 2.1 Å, which is a result of the very low count number as shown in the respective distance distribution in Figure 3.4(a). The ROTA potential for the atom types (GLU-OE1, LYS-NZ) shown in Figure 3.4(d) has been derived on a larger number of distance counts (167,029) then the potential for (HET-O.CO2, LYS-NZ), for which only 2,824 counts were available. Therefore the resulting potential for (GLU-OE1, LYS-NZ) is smoother than the potential for (HET-O.CO2, LYS-NZ).

There are generally no large differences between the potentials derived for docking into flexible proteins and those for side-chain prediction except that (i) the potentials for side-chain prediction are smoother due to the higher quality of the data set and (ii) these potentials have ten times more decoys in DLIB than in NLIB compared to potentials for docking which have an equal number of structures in NLIB and DLIB and that (iii) there were no potentials for interactions with ligand atoms derived for side-chain prediction. A comparative evaluation of the two versions of potentials for intra-protein interactions (see Figure 4.2 in Section 4.5.2) showed that the potentials derived for side-chain prediction are better suited for side-chain prediction than the potentials derived for docking. Therefore the current version of ROTA uses the potentials derived for docking for scoring of protein-ligand interactions and the potentials derived for side-chain prediction for scoring of intra-protein interactions.

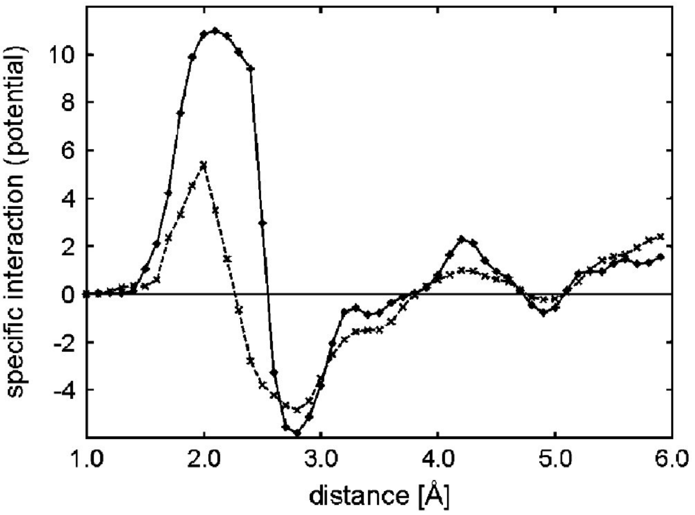
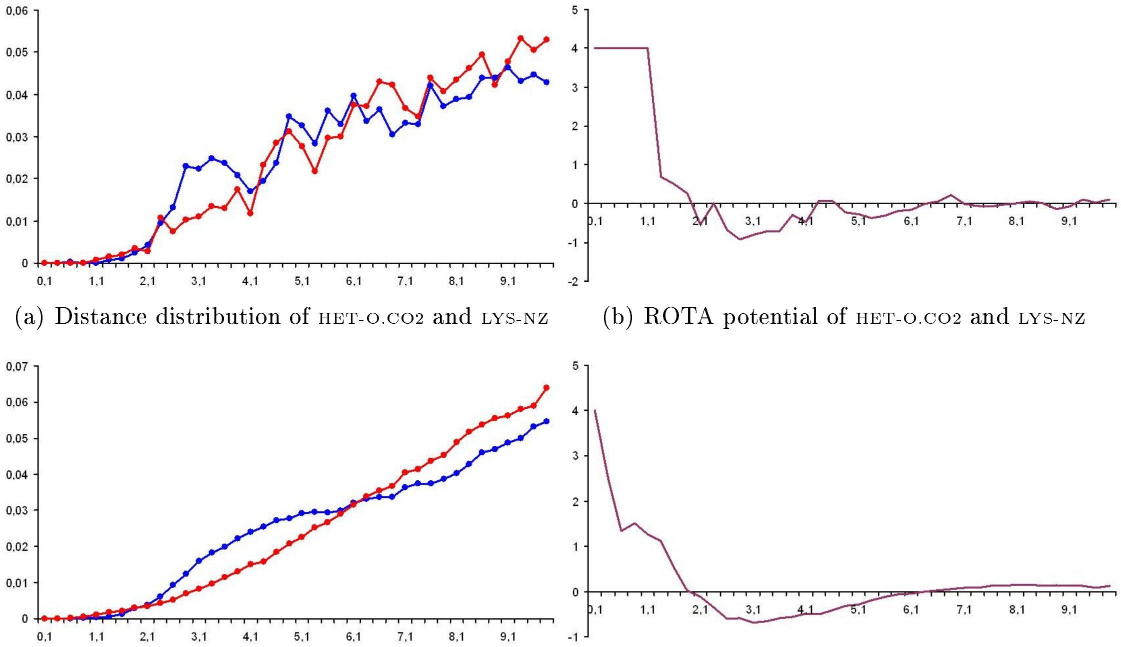


Figure 3.3: Distance-dependent pair potentials of DrugScore$_{\text{CSD}}$ (solid lines, boxes) and DrugScore$_{\text{PDB}}$ (dotted lines, crosses) for N.pl3 and O.co2 Sybyl atom types. From Velec et al. [214].



(a) Distance distribution of HET-O.CO2 and LYS-NZ

(b) ROTA potential of HET-O.CO2 and LYS-NZ

(c) Distance distribution of GLU-OE1 and LYS-NZ

(d) ROTA potential of GLU-OE1 and LYS-NZ

Figure 3.4: Distance distributions in the NLIB (blue) and DLIB (red) structure libraries (left side) and the respective ROTA potentials (right side). The x-axes represents distances in Angstroms, the y-axes probabilities (left, without unit) or ROTA scores (right, in kT units).

Table 3.3: Number of native structures at rank one among different decoy sets

Scoring function	4state	All-BBDep	All-Rot	Triple-BBDep	Triple-Rot
RAPDF	7/7	9/10	10/10	1/10	4/10
BBDep-Score	2/7	6/10	9/10	0/10	0/10
ROTA, All-BBDep	5/7	10/10	10/10	9/10	8/10
ROTA, All-Rot	6/7	10/10	10/10	7/10	9/10
ROTA, Triple-BBDep	3/7	10/10	10/10	6/10	7/10
ROTA, Triple-Rot	3/7	10/10	10/10	7/10	7/10

Table 3.4: Average Z-score of native structures among different decoy sets

Scoring function	4state	All-BBDep	All-Rot	Triple-BBDep	Triple-Rot
RAPDF	-3.189	-6.744	-9.160	-1.516	-1.385
BBDep-Score	-2.569	-3.190	-9.391	-0.298	-0.953
ROTA, All-BBDep	-3.196	-7.850	-9.026	-1.531	-1.413
ROTA, All-Rot	-3.547	-7.773	-9.391	-1.524	-1.453
ROTA, Triple-BBDep	-1.439	-6.965	-8.243	-1.328	-1.266
ROTA, Triple-Rot	-1.445	-6.975	-8.254	-1.333	-1.272

The small deviation between the Z-score found here for the RAPDF on the 4-state decoy set and that computed by Tosatto [209] is due to the fact that our scores do not include terms for intra-side-chain interactions.

3.6 Evaluation of ROTA: Identification of Native Side-Chain Conformations

All versions of the ROTA scoring function for side-chain prediction that were derived with the four different versions of the DLIB were tested for their ability to discriminate a native structure from a set of decoy structures.

This scenario is also known as a task for Model Quality Assessment Programs (MQAP), which are commonly applied for selecting the best model (or a subset of models) from a large pool of protein models that are generated with homology or *ab-initio* modeling methods (see Section 2.3). The purpose of this evaluation was to motivate the selection of a special version of the DLIB for generating ROTA potentials for side-chain prediction. Additionally, all structures were also scored with the scoring function RAPDF [184] and scores calculated by using the probabilities of the BBDep [43] like in SCWRL [26] with Equation 4.6 in Chapter 4.

The test set used in this evaluation is the MQAP test set together with the side-chain decoy sets generated for these structures (see Section 3.1) and the 4-state reduced [155] decoy set[4]. The 4-state reduced decoy set contains models with a variety of backbone

[4] downloaded from the *Decoys 'R' Us* website at http://dd.compbio.washington.edu

Table 3.5: Correlation coefficients between computed score and native-decoy RMSD in different decoy sets

Scoring function	4state	All-BBDep	All-Rot	Triple-BBDep	Triple-Rot
RAPDF	0.666	0.556	0.621	0.579	0.684
BBDep-Score	0.224	0.197	0.468	-0.042	0.271
ROTA, All-BBDep	0.572	0.571	0.582	0.639	0.714
ROTA, All-Rot	0.552	0.569	0.594	0.625	0.721
ROTA, Triple-BBDep	0.525	0.536	0.555	0.602	0.710
ROTA, Triple-Rot	0.528	0.536	0.556	0.601	0.711

conformations which is also reflected by the high RMSD values in Figure 3.1. All scores (also RAPDF scores) were calculated using a simple Python script that iterates over all protein atom pairs of a query model.

The results of this evaluation are summarized in Tables 3.3, 3.4 and 3.5. Table 3.3 shows the number of identified native structures in the decoy sets and Table 3.4 presents the Z-scores that were assigned to the native structures among the scores assigned to all other models of the decoy sets. Z-scores (also: standard scores) are obtained with $Z = \frac{x-\mu}{\sigma}$, where x denotes the raw scores, μ the mean of all scores and σ the standard deviation of all scores (considering all scores computed by a single scoring function on a single decoy set). The resulting Z-scores of the native structures show that the concept of using side-chain decoys for deriving the ROTA scoring function increases the ability of discriminating between native structures and decoy structures with disturbed side chains but also decreases the ability of discriminating between backbone decoys and the native structure: RAPDF can identify all native structures in the 4-state set (putting it to rank one), whereas all versions of the ROTA show a markedly decreased performance on this set.

The decoy sets Triple-BBDep and Triple-Rot are more challenging for all scoring functions, since they contain a larger fraction of decoys that are quite similar to the respective native structures (see Table 3.1). These sets are also handled quite well by all ROTA versions, since the rank of the native structure is one in more than 70% of all test sets. The RAPDF has a very low discriminatory power on these sets, identifying the native structure from the decoys in only 25% of all test cases. Also, the observed Z-scores show that the ROTA versions have a higher discriminatory power on the side-chain decoy sets than RAPDF, whereas the Z-scores are often quite similar. On the All-Rot test set, the Z-score achieved by the RAPDF is greater than that of three out of four ROTA versions. Nevertheless, the ROTA function derived from the All-Rot decoy set achieves greater or equal Z-score results than any of the non-ROTA scoring functions. The BBDEP scores show a low performance in the MQAP scenario in general, with a better performance on the decoys sets All-BBDep and All-Rot.

The correlation coefficients between the RMSD and score of the inspected structures generally support the above analysis, with two exceptions: first, RAPDF performs better than all other functions on the All-Rot decoy set. Second, in contrast to the prior analysis, the different versions of ROTA show nearly equal results. Based on this analysis, the ROTA version derived from the All-BBDep decoy set was chosen for later side-chain prediction with IRECS: it is likely that IRECS generates protein models whose failures are expected to be most similar to decoys from the set All-BBDep, since rotamers are selected from BBDep and their interaction with the local backbone are scored with the BBDep scores.

3.7 Evaluation of ROTA: Guiding a docking tool

The main purpose of ROTA is to guide the conformational sampling during side-chain prediction and ligand placement in the binding pocket. The ability of ROTA to support these applications is evaluated in Chapters 4 and 6. Here, the performance of ROTA is tested on a set of models of protein-ligand complexes and associated binding affinity data that was previously assembled by Wang et al. [223]. This data set consists of 100 protein-ligand complexes with experimentally determined binding affinity and a set of 100 alternative ligand poses for each of the 100 original complexes generated with AutoDock. This data set is one of the most widely used test sets for scoring functions for predicting ΔG_{bind}, and it therefore allows for comparing the performance of ROTA with that of other scoring functions. Wang et al. tested eleven scoring functions on this data set: LigScore [118], PLP [60], PMF [151, 150], LUDI-Score [12, 13, 14, 15], FlexX-Score [172], GOLD-Score [96, 97], DOCK-Score [49], ChemScore [44], AutoDock [147, 227, 228], DrugScore$_{\mathrm{PDB}}$ [63] and X-Score [222]. The scoring functions ITScore [86, 87], DOCK/FF [49, 142], DFIRE [238, 239, 240] and DrugScore$_{\mathrm{CSD}}$ [214] were later evaluated by their respective authors on the same data set. It should be mentioned that none of the previous authors discussed potential overlaps between this test set and protein structures that were used in training of one of their scoring functions. However, it must be assumed that for each of the complexes one or more identical or homologous proteins are contained in the ROTA structure libraries. A discussion of this topic can be found in Chapter 7.

3.7.1 Ranking of Ligand Poses

The ranking of ligand poses is the typical task of a scoring function that is part of a docking program. The goal is to identify a ligand pose with minimum RMSD from the native conformation of the ligand in the complex among a set of putative candidate conformations. The ligand poses that are used to test the scoring functions are usually generated by the docking program. Wang et al. generated the ligand poses with AutoDock, which uses a scoring function based on a force field [223]. Table 3.6 shows that the performance of ROTA is quite average compared to the other scoring functions. It is remarkable that ROTA and the other knowledge-based scoring functions DrugScore, DrugScore$_{\mathrm{CSD}}$, ITScore and PLP have a much better performance than the genuine scoring functions of the docking tools GOLD and DOCK which were designed especially for this task. Even the scoring function of AutoDock that was used to generate the ligand poses performs worse than ROTA. Among the scoring functions of docking programs the FlexX scoring function (F-Score) performs best. If the first two ranks are considered instead of only the first, the results of ROTA

Table 3.6: Percentages of native ligand poses ranked on top with different RMSD thresholds using docked ligand poses from the test set of Wang et al.

		Success rate [%]				
		RMSD	RMSD	RMSD	RMSD	RMSD
Scoring function	Source	≤ 1.0 Å	≤ 1.5 Å	≤ 2.0 Å	≤ 2.5 Å	≤ 3.0 Å
DrugScore$_{\text{CSD}}$	b	83	85	87	-	-
ITScore	c	72	79	82	85	88
Cerius2/PLP	a	63	69	76	79	80
SYBYL/F-Score	a	56	66	74	77	77
Cerius2/LigScore	a	64	68	74	75	76
DrugScore$_{\text{PDB}}$	a	63	68	72	74	74
Lennard Jones 12-6	b	65	66	68	-	-
Cerius2/LUDI	a	43	55	67	67	67
ROTA	e	52	59	66	67	69
X-Score	a	37	54	66	72	74
AutoDock	a	34	52	62	68	72
DFIRE	d	37	52	58	61	64
Cerius2/PMF	a	40	46	52	54	57
SYBYL/G-Score	a	24	32	42	49	56
SYBYL/ChemScore	a	12	26	35	37	40
SYBYL/D-Score	a	8	16	26	30	41

This table lists the percentages of protein targets for which the respective scoring functions could identify a near-native ligand pose using different RMSD cutoffs for ligand similarity. The sources of these values are: a) Wang et al. [223], b) Huang et al. [87], c) Zhang et al.[240], d) Velec et. al [214], e) this work.

increase significantly to 51 hits for 1.0 Å, 65 for 1.5 Å, 71 for 2.0 Å, 72 for 2.5 Å and 80 hits for 3.0 Å.

3.7.2 Prediction of Binding Affinities

An important task of docking programs is to estimate the binding free energy of a protein-ligand complex. The ability to predict the conformation of such a complex is a prerequisite of this task, but a scoring function is still required to estimate ΔG_{bind} from this conformation. It is a common technique to use one scoring function for prediction of the complex conformation and another scoring function for post-scoring of the final conformation [136]. In general, a scoring function for docking requires strong repulsive terms to prohibit the generation of ligand conformations with clashes, whereas post-scoring functions usually have much softer repulsive terms [200]. In this evaluation, ROTA was used to predict the contributions of the protein-ligand interactions to the binding free energy on the 100 X-ray structures of the test set of Wang et al. Although ROTA was derived to guide the ligand placement in FlexE, it is also quite suitable for predicting ΔG_{bind}.

Table 3.7: Spearman correlation coefficients between of binding scores and experimentally determined binding affinities of for ROTA and 15 other scoring functions on in Wang et al.'s test set.

Scoring function	Correlation coefficient	Source
ROTA	0.667	e
X-score	0.660	a
ITScore*	0.65	b
DFIRE	0.63	c
DrugScore$_{\text{CSD}}$	0.624	d
Cerius2/PLP	0.592	a
DrugScore$_{\text{PDB}}$	0.589	d
SYBYL/G-Score (GOLD)	0.569	a
SYBYL/D-Score (DOCK)	0.475	a
SYBYL/ChemScore	0.431	a
Cerius2/LUDI	0.430	a
DOCK/FF	0.40	b
Cerius2/PMF	0.369	a
Cerius2/LigScore	0.363	a
SYBYL/F-Score (FlexX)	0.283	a
AutoDock	0.141	a

The sources of these values are: a) Wang et al. [223], b) Huang et al. [87], c) Zhang et al. [240], d) Velec et al. [214], e) this work. *All authors present the Spearman correlation coefficient except Huang et al. who use the Pearson correlation coefficient. For ROTA the Pearson correlation coefficient to the measured binding affinity is 0.682.

This evaluation uses a correlation coefficient for determining the linear relationship between predicted and measured value pairs. In order to remain consistent with previous evaluations of other scoring functions, the Spearman correlation coefficient [78] is used here. It is defined as

$$R_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{3.7}$$

d_i denotes the difference in rank position of each pair member after creation of two different rankings based on the two pair member values. The two rankings are obtained by first sorting complexes by their binding affinity and second by sorting the complexes by the calculated scores. The Spearman correlation coefficient [78] is used for measuring the correlation between scores and binding affinities by Wang et al. and the respective authors for all scoring functions shown in Table 3.7, except ITScore, for which Huang et al. calculated the Pearson correlation coefficient. The correlation coefficient of ROTA is higher than reported previously for any other scoring function that was tested on this data set. The high performance of ROTA in predicting binding affinity is especially surprising as it was derived only from structure data without any training data for ΔG_{bind}. It can also be

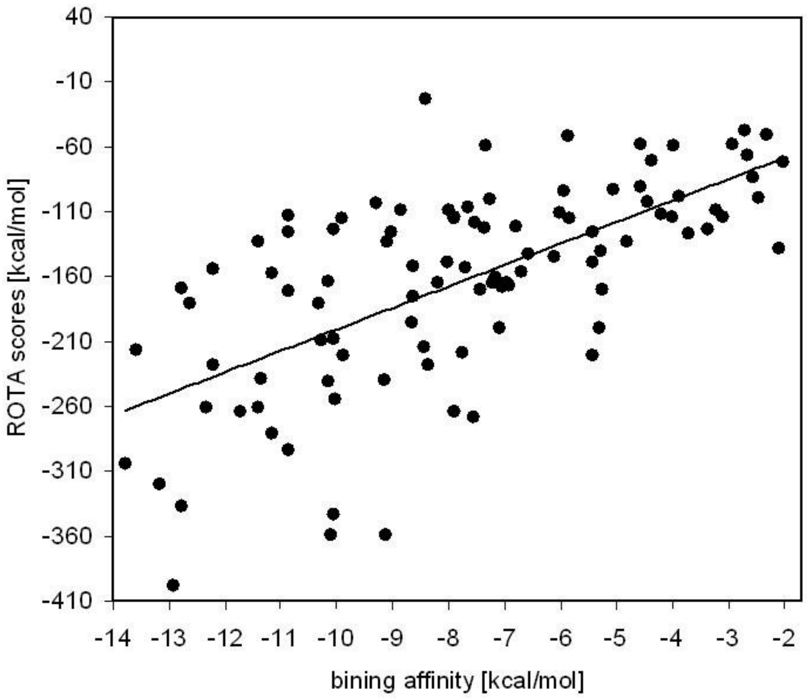


Figure 3.5: Binding affinity vs. ROTA scores

seen that most of the knowledge-based scoring functions also perform quite well on this task (ITScore, DFIRE, DrugScore$_{\text{CSD}}$, DrugScore$_{\text{PDB}}$), whereas most of the scoring functions based on force fields and empirical scoring functions perform worse. Since ROTA was not designed to score high quality structures, it was previously expected that the softness of the ROTA potentials would lead to a loss in sensitivity on this data set. Its good performance may be due to the large diversity of the structural data, which were derived by different crystallographers with different techniques using different chemical parameters and tolerance values for defining a protein structures as being correctly modeled. That ROTA can more easily tolerate these differences than any other scoring function may be the reason for it being more stable and reliable in its estimations.

Figure 3.5 shows a scatter plot of the calculated ROTA scores and the measured binding affinity. Since other authors plotted the prediction of their scoring functions in $\frac{kcal}{mol}$, this unit was also used in here instead of $\frac{kJ}{mol}$. Equation 2.2 was used with a temperature of $298K$ for conversion of the ROTA scores.

3.8 Discussion

In contrast to other scoring functions, two versions of ROTA have been derived for two different purposes. Each version required two different structure libraries, whereas other scoring functions usually require only one such library for derivation. The need for these four structure libraries arose from the main goal of ROTA: this is detecting those structural features that are highly discriminatory between correct and false protein structures. This aim distinguishes ROTA from many scoring functions for docking and molecular modeling, as these scoring functions usually aim at approximating contributions to the free energy of a molecular system. In order to find discriminative structural features, the range of correct and false protein structures must be efficiently and comprehensively sampled. These ranges depend on the particular modeling technique, and therefore it has been necessary to define structure libraries for the docking problem on the one hand and for the side-chain prediction problem on the other hand.

Although ROTA already achieves a remarkable scoring accuracy, it can be improved and extended through a number of approaches. First, the structure libraries can be extended. Since the size of the PDB grows exponentially with time and the quality of the deposited protein structures tends to increase, a regular update of the structural basis of ROTA is possible. This would allow for obtaining a more comprehensive description of the conformational space of protein-ligand complexes and the involved interactions, which in turn would allow for including more detail in the atom type schemes of ROTA. For example, the atom type scheme of ligands could be extended by charge information, which would improve the scoring of polar interactions. Also, special atom types for metal ions or the rare but important halogen atoms could be added. A second approach to extending the structure libraries would be to enhance the presented simulation procedure for the structure libraries with procedures that make use of other docking programs and rotamer libraries in order to generate additional conformations for the NLIB and DLIB structure libraries. This would potentially increase the performance when ROTA is used in concert with other docking and side-chain prediction programs, since this would allow for including modeling failures in the structure libraries that are characteristic of a variety of current modeling approaches, not just the failures that are characteristic of FlexX and side-chain prediction programs that use the BBDep. The most drastic change of ROTA and also the most promising one would be to define potentials not only for pairwise atom distances but also for more complex geometric features like special topologies of certain ligand subgroups. Summa et al. [202] already presented a knowledge-based scoring function that characterizes and scores the atomic environment of query atoms based on the occurrence of certain structural motifs composed of a small set of atoms. Their approach turned out to have higher performance than scoring functions like AMBER, CHARMM, DFIRE when discriminating protein decoy conformations from native conformations. These results indicate that scoring functions in general, and also ROTA, can profit from using more complex features like pairwise atom distances, if the underlying structural basis makes it possible to gather enough data for such larger features spaces.

Chapter 4

Flexible Side-Chain Prediction with IRECS

In this chapter the side-chain prediction program IRECS is presented. The name 'IRECS' stands for **I**terated **RE**striction of **C**onformational **S**pace which roughly summarizes the applied optimization procedure. Like other side-chain prediction programs (see Section 2.3.2) IRECS is able to predict single conformations for all side chains of a protein. Additionally, IRECS can also predict multiple conformations per side chain, where the number of conformations assigned to individual side chains depends on their respective flexibility. A first prototype version of IRECS was implemented in Python and later rewritten in C++ as a standalone program, with some I/O functions implemented in C for speed reasons. IRECS was published recently [72] and is accessible on the Internet[1]. The following section describes the probabilistic model that is used to characterize the problem of finding a meaningful representation of a protein model with user-defined flexibility of the side chains. Section 4.2 describes the IRECS algorithm which can compute heuristic solutions to this problem. Sections 4.3 to 4.6 evaluate and discuss the runtime and the performance of IRECS with respect to existing approaches and experimental results. A first application of IRECS is shown in Section 4.7 where IRECS is applied to an analysis of rotameric states of a single point mutant of the HCV protease NS3-4A. The last section discusses the general usefulness of IRECS and possible improvements.

4.1 Probabilistic Modeling of Side-Chain Flexibility

In this section, a probabilistic model for dealing with side-chain flexibility is introduced that is able to capture the most relevant aspects of side-chain flexibility by using discrete side-chain conformations. For docking, the most important characteristics of flexible side chains are first the possible conformations of a side chain and second the probabilities with which these conformations occur. For an efficient description of side-chain flexibility it is thus desirable to determine a set of discrete conformations that are suited best to represent the conformational space of a side chain, but the set should be limited in size. The aim of the

[1] http://irecs.bioinf.mpi-inf.mpg.de/

probabilistic model presented here is to provide an abstract formulation of the underlying optimization problem of this task.

The first definitions introduced here are required to formally describe the input and output of the IRECS algorithm. The first step in the side-chain prediction process is to generate initial coordinates of all side chains of the target protein. Next all assumed relevant conformations of all side chains of a protein are sampled by discrete rotations of all dihedral angles, e.g. from a rotamer library, of each side chain. Each side chain i of the protein has a set of rotamers L_i. At this step it is assumed that the sampling process is able to generate all relevant rotamers of side chain i and that all of them are contained in the set L_i. Let X_i be the name of an arbitrary subset of L_i. The list $L = L_1, ..., L_i, ...L_n$ is called the rotamer space of the protein. The list $X = X_1, ..., X_i, ...X_n$ is analogously called a reduced rotamer space of the protein. Each solution candidate of the IRECS algorithm consists of the protein backbone and a certain reduced rotamer space. The next definitions are used to characterize such a solution candidate. From the two rotamer spaces L_i and X_i different conformations of the protein can be generated by selecting one conformation from L_i or X_i, respectively, for side chain i. $G(L)$ is the set of conformations that can be constructed in this way by taking side-chain conformations from L and analogously for $G(X)$.

A good rotamer reduction contains not only a large number of possible protein conformations, but also the most probable protein conformations: given a scoring function that can approximate the energy of a conformation $E(C)$, each conformation C in $G(X)$ is assigned a Boltzmann probability $P(C)$ that depends on this energy $E(C)$ and the energetic states of all other conformations $\hat{C}$ in $G(L)$:

$$P(C) = \frac{e^{\frac{-E(C)}{kT}}}{\sum\limits_{\hat{C} \in G(L)} e^{\frac{-E(\hat{C})}{kT}}} \tag{4.1}$$

k is the Boltzmann constant and T is the temperature of the molecular system. For a given reduction X of L, the probability $P_G(X)$ denotes the sum of all Boltzmann probabilities of all conformations in X:

$$P_G(X) = \sum_{C \in X} P(C) = \frac{\sum\limits_{C \in G(X)} e^{\frac{-E(C)}{kT}}}{\sum\limits_{\hat{C} \in G(L)} e^{\frac{-E(\hat{C})}{kT}}} \tag{4.2}$$

An optimal rotamer reduction X maximizes $P_G(X)$. Such a reduction is the optimal choice of rotamers for describing the conformational space accessible to the side chains of a protein. Of course, this can be done trivially by removing no conformation, so that $X = L$. To make the problem formulation more useful, the sizes of the rotamer sets are limited by some flexible measure. Such a constraint can pose an upper bound of the average number of rotamers per side chain (called the rotamer density $\rho_{rot}(X)$ of a reduction X:

$$\rho_{rot}(X) = \frac{1}{n} \sum_{i=1,\ldots,n} |X_i| \tag{4.3}$$

Now one can formulate the flexible side-chain prediction problem by using reduced rotamer spaces X with bounded rotamer density. The optimal rotamer reduction has a maximum cumulative Boltzmann probability and a rotamer density that is bounded from above by a number b. The flexible side-chain prediction problem is therefore to find a rotamer reduction X among all possible rotamer reductions Y that fulfill the following criteria:

$$\forall\, Y : \rho_{rot}(Y) \leq b \;\Rightarrow\; P_G(X) \geq P_G(Y) \tag{4.4}$$

A reduction X that meets this constraint enables the generation of a set of protein conformations $G(X)$, which altogether have a maximal sum of Boltzmann probabilities compared to other candidate reductions Y with rotamer density below b.

The rigid side-chain problem is a special case of the flexible side-chain prediction problem formulated above if $b = 1$. In the rigid case, the global minimum energy conformation (GMEC) of a protein is searched for using a fixed backbone conformation and multiple side-chain conformations. As Pierce and Winfree showed [160], the rigid side-chain problem is NP-hard if pairwise potentials are used for scoring energy (see Section 2.3.2). A proof for NP-completeness of the flexible side-chain prediction problem was not attempted here, but the conversion made above should illustrate the hardness of this problem and the need for approximate solutions.

4.2 The Optimization Algorithm IRECS

The aim of the IRECS algorithm is to find an approximate solution for the flexible side-chain prediction problem. IRECS must work heuristically because the space $G(X)$ is too large for explicit enumeration. IRECS follows a simulated annealing strategy [31, 104] that allows transition from high energy states to low energy states and vice versa while ensuring that 'downhill' moves to low energy states are more frequent. The energy of a state is evaluated by using the effective energy approach that is carried over from mean field theory as described by Koehl and Delarue [110, 112]. The overall strategy of IRECS is to remove rotamers one-by-one until each side chain is assigned a single rotamer or until the user decides to interrupt the optimization process. In each step of the optimization process all rotamers are scored and a single rotamer is chosen for removal that is supposed to contribute least to the conformational space of the protein.

4.2.1 Sampling the Full Rotamer Space

Given a protein model with a rigid backbone conformation, IRECS ignores all previous side-chain conformations and builds initial side-chain conformations using the definitions for bond lengths and angles in the topology file of the CHARMM [25, 137] force field[2]. By default, IRECS uses the residue sequence of this backbone as the target sequence of the

[2] http://www.charmm.org/

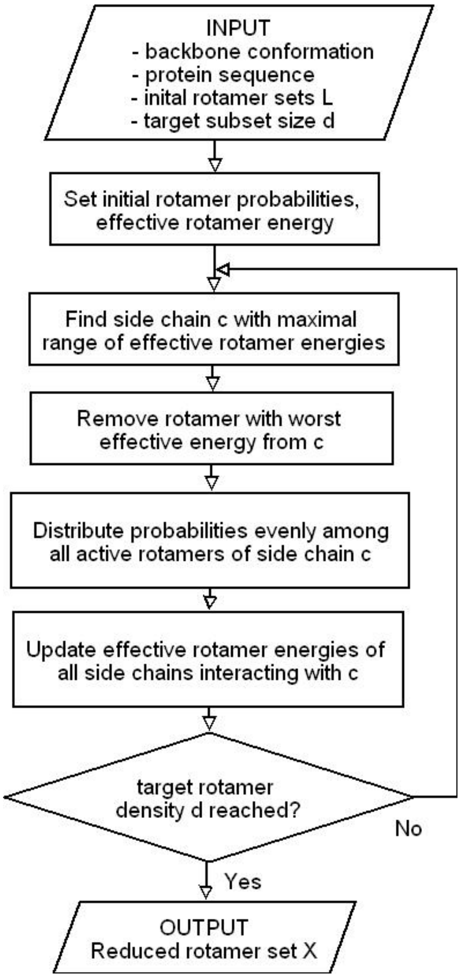

INPUT
- backbone conformation
- protein sequence
- inital rotamer sets L
- target subset size d
Set initial rotamer probabilities, effective rotamer energy
Find side chain c with maximal range of effective rotamer energies
Remove rotamer with worst effective energy from c
Distribute probabilities evenly among all active rotamers of side chain c
Update effective rotamer energies of all side chains interacting with c
target rotamer density d reached?
No
Yes
OUTPUT
Reduced rotamer set X

Figure 4.1: Flowchart of the IRECS algorithm

final model. Optionally, IRECS can read in a target sequence from a file in FASTA format. IRECS then uses the provided protein model as a template and aligns the target sequence to the sequence of the template backbone. Side chains of the template protein are then changed to the side chains of the respective aligned target residues.

The conformational space of each side chain is then represented by an ensemble of rotamers: all rotamers that are defined for the respective amino acid in the BBDep are sampled with repeated rotations of the initial side-chain conformation. Each rotamer of the ensemble is also assigned a probability that measures its influence on the other rotamers: if the probability is larger than zero the rotamer is called active and its interactions are used to calculate the energy state of the system. If this probability is set to zero, the rotamer is effectively removed from the ensemble and its interactions are neglected during all subsequent computations. At the beginning the probabilities of an ensemble are distributed uniformly among all its rotamers.

4.2.2 Side-Chain Interactions

IRECS knows three different interaction partners: the fixed protein backbone, a set of ligands with fixed conformation and rotamers of other side chains. Since the conformational space of the side chains is known beforehand, it is possible to compute all relevant interactions before the optimization procedure starts and to store them in a matrix for fast lookup during the optimization. All relevant interactions are scored with ROTA potentials (see Equation 3.5). Here, the term U_{inter} is applied to all interaction partners y of a certain side-chain conformation x and is calculated by iteration over all $n \times m$ atom pairs of the respective interaction partners.

$$U_{\text{inter}}(x, y) = \sum_{i}^{n} \sum_{j}^{m} \Delta\Delta S_{ij}^{ROTA}(d) \tag{4.5}$$

In this and the following chapter, the potentials derived for side-chain prediction are used if not otherwise stated. The interactions of rotamers with the backbone are split into interactions with the local backbone part and far interactions with backbone atoms that do not belong to the residue of the particular side chain. The interactions with the local backbone part are denoted by U_{self}. Their contributions are weighted with the rotamer probabilities defined in the BBDep and therefore depend on the local backbone conformation that is characterized by Φ and Ψ (see Section 2.3.2). The functional form of this term is the same as for the side-chain prediction tool SCWRL [26]:

$$U_{\text{self}}(x_i) = -2.5 \cdot \ log \frac{p(x_i|\Phi_i, \Psi_i)}{\max\limits_{y_i \in C_i} p(y_i|\Phi_i, \Psi_i)} \tag{4.6}$$

The scaling factor -2.5 is required for adding U_{self} to U_{inter}. It was found empirical on a training set of 34 X-ray structures (see Appendix C.1) and by minimizing the side-chain RMS deviation between predicted models (single side chains on a fixed backbone) and the respective X-ray structures.

4.2.3 The Effective Energy Approach

IRECS requires an energy score for each active rotamer for assessing its value for the conformational space of the protein. As proposed by Koehl and Delarue the energy contribution of a side-chain conformation can be approximated with the effective energy approach (see Section 2.3.2). If there are multiple side-chain conformations in a system, each single conformation contributes to the system energy by all energy contributions from interactions with all conformations of other side chains, weighted by the probabilities of these conformations.

$$E_{\text{eff}}(x_i) = U_{\text{self}}(x_i) + U_{\text{inter}}(x_i, b) + \sum_{\substack{j=1 \\ j \neq i}}^{s} \sum_{y_i}^{C_j} p(y_j) U_{\text{inter}}(x_i, y_j) \tag{4.7}$$

At the start of the optimization this approximation is quite bad since the initial size of the ensembles depends on the unadjusted number of rotamers that are defined in the

BBDep for the respective amino acids. The energy contributions of interacting side chains are therefore dominated by many clashing rotamers. With decreasing rotamer density, the conformational spaces of the side chains are adjusted to their environment and clashes are removed, so that the energy approximation becomes more accurate.

4.2.4 Minimizing the Effective Energy

IRECS now carries out an optimization procedure that reduces the number of rotamers of the proteins (and therefore its flexibility) and also lowers the system energy. IRECS removes a single unfavorable rotamer in each step and thus moves slowly but continuously to a minimum energy conformation. This procedure either stops if each side chain has only one rotamer left or if the user interrupts the optimization. IRECS allows to define a threshold for the minimum rotamer density of the protein model and if this is reached the optimization ends.

Reducing the Conformational Ensembles

In this step IRECS searches all side chains for a rotamer that has the lowest chance to contribute to the conformational space of its side chain. This is a rotamer has the highest effective energy among all rotamers of the same side chain. As in the SCMF approach [110] the Boltzmann probability of a rotamer in its ensemble can be calculated based on the effective energy of the rotamers in the ensemble:

$$p(x_i) = \frac{e^{\frac{-E_{\text{eff}}(x_i)}{kT}}}{\sum_{y_i}^{C_i} e^{\frac{-E_{\text{eff}}(y_i)}{kT}}} \tag{4.8}$$

A test that searched for rotamers with minimum Boltzmann probabilities according to Equation 4.8 and iteratively deleted them led to improper optimization behavior. As previously noted, the effective energies of many rotamers of the protein are quite unrealistic at the start and therefore these side chains possess unrealistic Boltzmann probabilities which can describe the observed behavior. Therefore, a more robust selection criterion was developed to find the side chain which should lose one rotamer. The Boltzmann sum of Equation 4.8 is dominated by the few rotamers which have a very low effective energy, and thus, a rotamer with a low Boltzmann probability can be found at a side chain s which has a large energy range ΔE_{eff} between the best and worst rotamer x_i and y_i:

$$\Delta E_{\text{eff}}(s) = \max_{y_i \in C_i} E_{\text{eff}}(y_i) - \min_{x_i \in C_i} E_{\text{eff}}(x_i) \tag{4.9}$$

The side chain with the largest ΔE_{eff} is the one which loses one rotamer. The rotamer with greatest effective energy of this side chain is removed by setting its probability to zero. The probabilities of the remaining rotamers are distributed uniformly among the rotamer ensemble so that they again sum up to one. It is notable that the flexibility of a side chain depends on the ability to easily change its rotameric state which in turn requires the

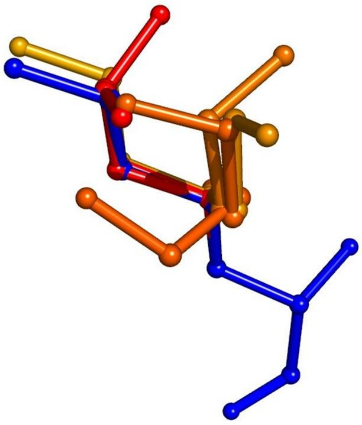


Figure 4.2: Side chain with multiple rotamers colored by Boltzmann probabilities (blue = X-ray structure, red = low probability, yellow = high probability)

rotamers to have similar energetic states. Equation 4.9 therefore tends to select inflexible side chains and their energy range decreases through the removal of the worst rotamer.

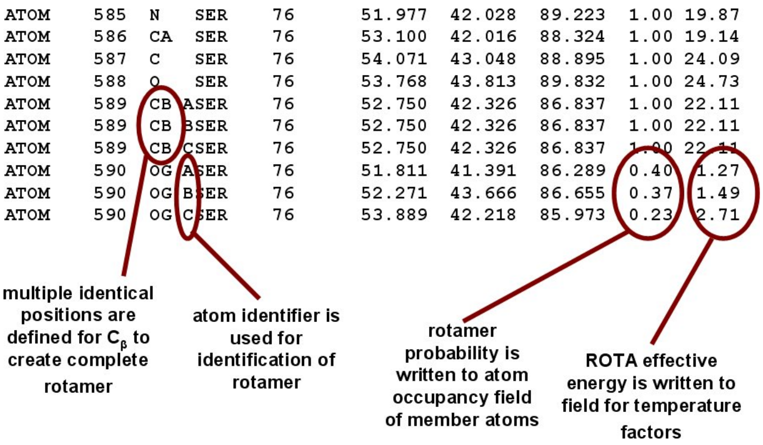
Adjusting the Effective Energy

After the probabilities of the rotamers of the selected side chain have been changed, the effective energies of the interacting side chains must be recomputed. Specifically the interactions to the removed rotamer are deleted and the interactions to the remaining rotamers become stronger, since the probabilities of these rotamers increased by a small amount. This step is the most resource-intensive step of the IRECS optimization procedure because of the many pair interactions between rotamers.

4.2.5 Output of the Final Model

After IRECS has finished the optimization process, the remaining rotamers represent the selected conformational space of the protein side chains. This reduced space is now compared to the full rotamer space. All rotamers of the full rotamer space are assigned effective energies using the post-optimization rotamer probabilities. The Boltzmann law for rotamers in Equation 4.8 is then applied to all rotamers of the full rotamer space of each side chain. The rotamer probabilities of rotamers from the full rotamer space now sum up to one. The sum of probabilities from all selected rotamers of a side chain can therefore be used for evaluating the ability of IRECS to find a representative subset of rotamers for each side chain. Also, the rotamer probabilities quantify the preference of the side chain to adopt a conformation using only the selected rotamers. An example is shown in Figure 4.2. The figure depicts the rotamers of a single side chain, which are colored according to their Boltzmann probabilities from yellow (high probability) to red (low probability). The native conformation of the side chain is drawn in blue.

The protein model is then written to a PDB file (see Figure 4.3 for a short example). The alternative side-chain conformations are written as multiple atom insertions for the



```
ATOM     585  N    SER    76       51.977  42.028  89.223  1.00 19.87
ATOM     586  CA   SER    76       53.100  42.016  88.324  1.00 19.14
ATOM     587  C    SER    76       54.071  43.048  88.895  1.00 24.09
ATOM     588  O    SER    76       53.768  43.813  89.832  1.00 24.73
ATOM     589  CB  ASER    76       52.750  42.326  86.837  1.00 22.11
ATOM     589  CB  BSER    76       52.750  42.326  86.837  1.00 22.11
ATOM     589  CB  CSER    76       52.750  42.326  86.837  1.00 22.11
ATOM     590  OG  ASER    76       51.811  41.391  86.289  0.40  1.27
ATOM     590  OG  BSER    76       52.271  43.666  86.655  0.37  1.49
ATOM     590  OG  CSER    76       53.889  42.218  85.973  0.23  2.71
```

multiple identical positions are defined for C_β to create complete rotamer

atom identifier is used for identification of rotamer

rotamer probability is written to atom occupancy field of member atoms

ROTA effective energy is written to field for temperature factors

Figure 4.3: Excerpt from a PDB file for a residue modeled with IRECS. Three rotamers were assigned by IRECS to this residue. The nearly equal distribution of rotamer probabilities suggests that the rotamer states can be easily exchanged, which results in a highly flexible model of this side chain.

respective side chains. The Boltzmann probabilities are given in the field for atom occupancy, and the temperature field contains the effective energy of the rotamer to which the respective atom coordinate belongs.

4.3 Runtime

The runtime of the IRECS algorithm primarily depends on the number of residues of the target protein and the average number of rotamers per side chain at the beginning. Let n be the average starting number of rotamers per side chain and let p be the number of side chains to be predicted, then a single IRECS iteration requires time $O(n^2p)$: after removal of a rotamer from side chain s, $O(np)$ other rotamers can interact with the $O(n)$ rotamers of s and need their effective rotamer energies updated. If the minimization is continued until rotamer density one, $O(np)$ IRECS iterations must be performed. The time complexity of a complete IRECS optimization is therefore $O(n^3p^2)$. This is close to the complexity of the initial (and inevitable) computation of all pairwise (atomic) rotamer-rotamer interactions ($O(n^2p^2a^2)$, where a is the average number of heavy atoms per rotamer). Actually, both tasks usually require about the same amount of runtime (seconds to a minute) in real-life scenarios.

4.4 Measures for Comparison of Conformations and Evaluation of Prediction Accuracy

The following measures are common in the field for comparing conformations of proteins, protein fragments (e.g. side chains) and ligand positions, for characterizing protein properties and for estimating prediction accuracy. A definition is given here since these measures are applied in the following evaluations and details of these measures are usually different between authors, e.g. if hydrogen atoms are considered or which specific cutoffs are chosen.

4.4.1 Root Mean Square Deviation

The root mean square deviation (RMSD) is a measure for structural similarity if applied to two equal sized sets A and B of Cartesian atom coordinates. d is the Euclidean distance of two atoms.

$$\mathrm{RMSD}(A, B) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} d(A_i, B_i)^2} \tag{4.10}$$

It is common to compare only the positions of heavy atoms since the positions of hydrogen atoms are rarely resolved by X-ray crystallography. They are always bound to a single heavy atom and in many cases their position is determined by the position of this atom. All RMSD values presented in this work follow this convention. The RMSD is used here for benchmarking the performance of docking tools and side-chain prediction tools. Whenever an RMSD value is given for a computed ligand conformation or a set of such conformations, this value was computed between atoms of this conformation and the corresponding atoms in a reference conformation. If the RMSD of side chains is computed, only the heavy atoms of the side chains are considered, usually excluding the C_β atom, as its position is mainly determined by the backbone, which is held fixed during side-chain prediction. The side chains of arginine, aspartic acid, glutamic acid, phenylalanine, and tyrosine have symmetric conformations. Therefore the RMSD of two symmetric side chains is always calculated for all symmetric alternatives of atom matches between the chains and the minimum RMSD over all alternative matches is used. If multiple rotamers from a side-chain ensemble are compared to a single reference rotamer then the RMSD to the best matching rotamer (yielding the minimum RMSD) in the ensemble is used.

4.4.2 Chi-Match Criterion

Two side chain conformations are compared with each other by comparing all dihedral angles of the side chains with each other – or – just those dihedral angles noted as indices i of χ_i. Two dihedral angles are regarded as matched if their difference falls below a given cutoff. Since this work builds on the coarse-grained BBDep, the cutoff value used here is 40°. Lower cutoff values are usually applied if finer rotamer libraries are used for side-chain sampling [141, 233]. Two side-chain conformations can partially match up to a certain dihedral angle. Such a partial match is denoted with χ_1 for a match of the first dihedral angle, $\chi_{1,2}$ for the first two dihedral angles, $\chi_{1,2,3}$ for the first three dihedral angles and $\chi_{1,2,3,4}$ for a match of all four dihedral angles. The advantage of this measure over the RMSD is that the backbone conformations of two side chains need not be superimposed with each other before calculation.

4.4.3 Relative Solvent Accessibility

Usually the position of a side chain is much harder to predict if the side chain is at the surface of the protein, since it can interact with the solvent, which just slightly restricts its flexibility. In contrast, the side chains in the protein core are much less flexible because they are packed very tightly. Figure 2.3(a) shows the solvent-excluded surface of a protein that was calculated with the 'rolling ball algorithm' [177]. A common approach to distinguishing surface residues from core residues is to calculate the accessible surface area of each residue [126]. This is done here with the program Naccess[3] (version 2.1.1) [89], which calculates the relative solvent accessibility (RSA) for each side chain. Any side chain with an RSA below 20% is considered as belonging to the protein core in this work.

4.5 Accuracy of IRECS when Predicting Rigid Side-Chain Conformations

Since the rigid side-chain problem is a special case of the flexible side-chain prediction problem an evaluation is performed that measures the accuracy of IRECS o predict protein models with single side-chain conformations. A test set of proteins with high-quality X-ray structures was therefore assembled from the PDB. The PISCES server[4] [219, 220] was used to filter the PDB for structures that had a resolution better than 1.5 Å, R and R_{free} below 0.3, a pairwise sequence identity below 25% and only single conformations for all side chains. From the resulting 194 protein structures 34 were chosen randomly for the training (see Appendix C.1) of the weighting factor for the scoring functions (see Eq. 4.6), and 160 for the test set (see Appendix C.2).

4.5.1 Matching Dihedral Angles of X-Ray Structures

IRECS was used to build protein models with a single conformation per side chain. Figure 4.4 shows the average prediction accuracy of IRECS for up to four dihedral angles for each amino acid with a flexible side chain. The achieved prediction accuracy is not much different (>10% increase or decrease) from the average prediction accuracy reported for other methods [129, 159]. The low accuracy of predicting the χ_4 dihedral angle of arginine and lysine is due to the high flexibility of these side chains and the large number of rotamers that are required to represent their conformational space. Amino acids with large ring systems like histidine, phenylalanine and tryptophane have an exceptional high prediction accuracy also for their χ_2 dihedral angle which can be explained by the hydrophobic nature of these side chains and their relative accumulation in the protein core, where side chains are much easier to predict.

In a second test the loss of native-like conformations during the IRECS optimization was analyzed. In the test the rotamer ensembles during the IRECS optimization were compared with the rigid conformations of the respective X-ray structures from the test set to inspect when these optimal rotamers get lost. A rotamer was defined as near-native if the χ_1 angle

[3] http://wolf.bi.umist.ac.uk/naccess/

[4] http://dunbrack.fccc.edu/PISCES.php

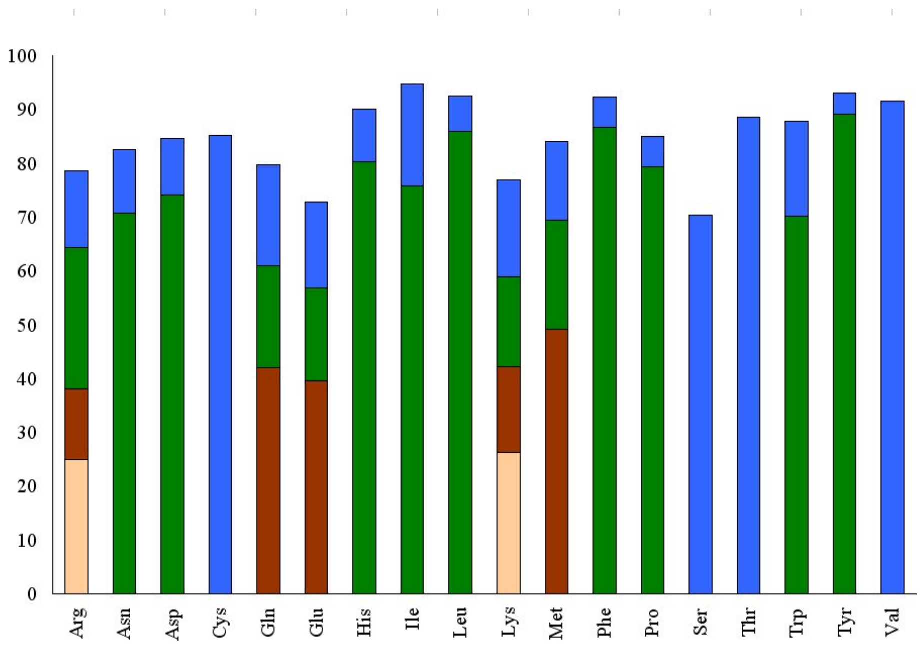


Figure 4.4: Prediction accuracy for side chains of different amino acids. The bars represent the percentages of successful χ-match tests for all dihedral angles (blue: χ_1, green: $\chi_{1,2}$, red: $\chi_{1,2,3}$, yellow: $\chi_{1,2,3,4}$) of the side chains.

matched that of the respective side chain in an X-ray structure. Figure 4.5 shows the average percentage of rotamer ensembles fulfilling this criterion for different stages of the rotamer removal process, measured by the current overall density of rotamers ρ_{rot}. Each of these curves describe this average percentage, once using the IRECS scoring function (ROTA + BBDep), using both components of the scoring function, or using randomized rotamer interaction scores.

IRECS performs quite well using ROTA and BBDep independently on this test. In the range of $3 \geq \rho_{rot} \geq 1$, the performance of IRECS using BBDep drops significantly, whereas IRECS using ROTA has a rather stable performance also in this region. When IRECS uses both the ROTA and BBDep components of its scoring function, the performance is highest for all levels of rotamer density. However, the average percentage of side chains without a rotamer having a correct χ_1 angle in the ensemble more than doubles in the last steps of the algorithm, increasing from 6% to 15% for $2 \geq \rho_{rot} \geq 1$. This high risk of removing the correct rotamer in the last steps of the algorithm has three reasons. First, at the start of the optimization the rotamer ensembles contain many rotamers that can be clearly identified as useless because of their high effective rotamer energy. In later iteration steps, the removal criterion of equation 4.9 causes all rotamers of an ensemble to have almost the same effective energy (and also rotamer probabilities) and this makes it quite hard to identify the correct rotamer. Second, the inaccuracy of the scoring function leads to incorrect removals of rotamers. Specifically, towards the end of the IRECS algorithm, differences in the effective energies may drop below the noise level that is presented by the inherent inaccuracy of the scoring function. Last, this behavior can be caused by the

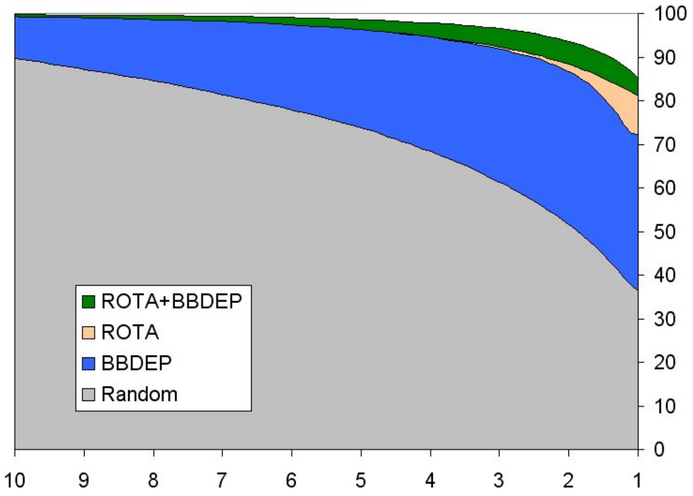


Figure 4.5: Average prediction accuracy of IRECS for different rotamer densities and scoring functions

flexibility of some side chains that require a large ensemble of rotamers for an adequate representation of their conformational space. Since the performance of IRECS is evaluated with only single conformations for all side chains here, the chance of selecting the same conformation as observed in the X-ray structure is quite low for these side chains. This also shows that reducing the rotamer density beyond the natural conformational limit of certain side chains helps little in raising the accuracy of the generated protein conformation but worsens it by neglecting any conformational side-chain flexibility.

4.5.2 Comparison with Other Tools for Side-Chain Prediction

IRECS can directly be compared to other side-chain prediction tools if IRECS is configured to generate models with single side-chain conformations. The tools SCWRL [26] and SCAP [233] were selected for comparison because of their popularity, availability, and high accuracy. They are also of interest for a comparison since they use fundamentally different optimization algorithms and scoring functions. For this test all three tools use only the backbone conformation of the protein structures as input and are given no conformational information of the side chains. However, such information is used indirectly, since some of the proteins are part of the training sets of the BBDep (used by SCWRL and IRECS), ROTA and most likely also the CHARMM [25] and AMBER force fields [157], which are used by SCAP. The performance of all three tools is measured by the similarity of the generated models to the original X-ray structures. For comparison of side-chain conformations the side chain RMSD (see Section 4.4.1) and the χ-match criterion (see Section 4.4.2) is used. All tools are tested sequentially on the same machine (AMD Opteron V20z).

Table 4.1 shows the averaged results for the whole test set, once for all side chains and once for all side chains belonging to the core. In general, all programs perform better for core residues whereas the residues at the surface are much harder to predict, as it was previously observed [123]. This can be explained by three main effects, (i) higher side-chain flexibility at

Table 4.1: Comparison of side-chain prediction performance on the target backbone

All side chains

	Overall RMSD [Å]	Average RMSD [Å]	χ_1 [%]	$\chi_{1,2}$ [%]	Runtime (sec)
SCWRL	1.677	0.849	82.3	68.0	4.7
SCAP	1.605	0.823	84.0	70.6	161.0
IRECS	1.551	0.775	84.7	71.6	23.1

Only *core* side chains

	Overall RMSD [Å]	Average RMSD [Å]	χ_1 [%]	$\chi_{1,2}$ [%]
SCWRL	1.191	0.572	90.2	80.1
SCAP	1.120	0.529	91.6	83.9
IRECS	1.046	0.502	92.0	82.2

The measures *overall RMSD* and *χ-match test* are first measured for each individual protein of the 160 proteins in the test set and then averaged among all these proteins. The *average RMSD* is measured as an average over all side chains in the test set, regardless to which protein structure they belong. The runtime is averaged without the runtime on PDB code 1gd0, since SCWRL required more than 16 hours for it.

the surface, (ii) (not modeled) interactions with the solvent and (iii) crystal packing effects (see 2.1.3). Although IRECS has small advantages in most of the similarity measures, the accuracy of all three tools is quite similar. This is especially interesting considering the large differences of the applied scoring functions (SCWRL: steric clash plus rotamer probabilities, SCAP: force field, IRECS: ROTA plus rotamer probabilities). These differences however become apparent when looking at the runtime: although SCWRL determines an optimal rotamer selection according to its model of configuration space, it is the fastest, whereas IRECS and SCAP use heuristic algorithms and are much slower in all but one case. This shows that the runtime of the programs is largely determined by the size of the rotamer libraries (IRECS and SCWRL use BBDep, SCAP a fine coordinate rotamer library) and the distance cutoffs of the scoring functions (SCWRL: 3.4 Å, IRECS: 10 Å, SCAP: unknown, but 12 Å - 14 Å are typical cutoffs for CHARMM), which determine the number of atom pairs that must be scored.

Since it is hard to assess the respective influence of the optimization algorithm or the scoring function on the prediction accuracy, four different experimental implementations of IRECS with different versions of the scoring function and simplifications for scoring were tested on the same data set. Table 4.2 shows the different results for these implementations. The IRECS implementation ROTA$_{\text{SCP}}$ uses the potentials derived for side-chain prediction, which is the default setting for IRECS. ROTA$_{\text{DFP}}$ uses the potentials derived for docking into flexible proteins, which were derived on a different set of X-ray structures and uses a different set of atom types (see Sections 3.2 and 3.3.2). As expected the performance of

Table 4.2: Accuracy of IRECS using different ROTA potentials and simplifications

All side chains

	Overall RMSD [Å]	Average RMSD [Å]	χ_1 [%]	$\chi_{1,2}$ [%]
ROTA$_{\mathrm{SCP}}$	1.551	0.775	84.7	71.6
ROTA$_{\mathrm{DFP}}$	1.812	1.053	83.8	69.3
OPTIMA$_{\mathrm{crystal}}$	1.500	0.753	85.4	72.9
OPTIMA$_{\mathrm{rotamer}}$	1.707	0.957	85.8	72.1

Only *core* side chains

	Overall RMSD [Å]	Average RMSD [Å]	χ_1 [%]	$\chi_{1,2}$ [%]
ROTA$_{\mathrm{SCP}}$	1.046	0.502	92.0	82.2
ROTA$_{\mathrm{DFP}}$	1.099	0.680	91.4	80.7
OPTIMA$_{\mathrm{crystal}}$	0.999	0.483	91.9	82.8
OPTIMA$_{\mathrm{rotamer}}$	1.019	0.581	93.3	83.0

The IRECS implementation ROTA$_{\mathrm{SCP}}$ uses the potentials derived for side-chain prediction (default, same values as in Table 4.1), ROTA$_{\mathrm{DFP}}$ uses the potentials derived for docking into flexible proteins. OPTIMA$_{\mathrm{crystal}}$ predicts each side chain with rotamer interactions from the X-ray structure. OPTIMA$_{\mathrm{rotamer}}$ predicts each side chain with rotamer interactions with best matching rotamers from BBDep to the respective side chain in the X-ray structure.

ROTA$_{\mathrm{DFP}}$ is worse than the performance of ROTA$_{\mathrm{SCP}}$ on this test. All further tests of IRECS were therefore performed using ROTA$_{\mathrm{SCP}}$.

In order to clarify if the scoring function or the IRECS algorithm is to blame for the suboptimal accuracy of IRECS (84.7% χ_1 accuracy means that one out of six side-chain conformations is completely wrong) a simulation was performed that aimed at eliminating the influence of the IRECS algorithm and at measuring the performance of the scoring function. In this simulation the prediction of a side chain is simplified by choosing the native conformation for all other side chains of the protein during scoring instead of a rotamer ensemble. Two setups of this simulation were chosen: once, the native conformation of the surrounding side chains was taken directly from the crystal structure (OPTIMA$_{\mathrm{crystal}}$), and once, the best matching rotamer from the BBDep to the side-chain conformation of the X-ray structure was chosen (OPTIMA$_{\mathrm{rotamer}}$). These different setups can help to measure the performance decrease that results from using the BBDep instead of a much finer rotamer library. When comparing ROTA$_{\mathrm{SCP}}$ with OPTIMA$_{\mathrm{rotamer}}$ one can see that using the optimal reference environment for rotamer scoring is just slightly better than using the effective energy approach of IRECS. This witnesses that the IRECS algorithm has a good approximation ability and that there is only a small loss in performance that results from using a heuristic approach.

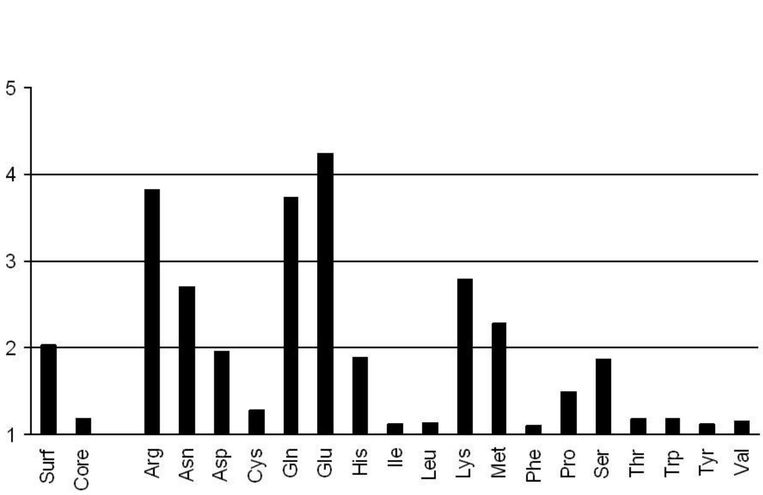


Figure 4.6: Average number of rotamers assigned to different amino acid types

The performances of the two simulation setups OPTIMA$_{\text{crystal}}$ and OPTIMA$_{\text{rotamer}}$ are generally quite similar. The RMSD measures of OPTIMA$_{\text{crystal}}$ are better than those of OPTIMA$_{\text{rotamer}}$, whereas the situation is reversed for the χ_1 and $\chi_{1,2}$-match tests. Such contradictory results can be seen occasionally (also in Table 4.1 for IRECS and SCAP core residues) and can be explained by the fact that RMSD values for different amino acids profit from correct dihedral angles to different amounts. A false χ_1 of tryptophan results in a very bad RMSD of the model, whereas a false χ_1 of serine just worsens the RMSD only slightly. The results show that the fixed rotamer approach of the BBDep does not lower the accuracy of IRECS significantly and therefore the BBDep is an appropriate choice for initial definition of the full rotamer ensembles. These observations imply that either the scoring function, the modeling of the protein environment (no modeling of explicit solvent, cofactors or binding partners) or the natural side-chain flexibility circumvent the correct prediction of side-chain conformations in X-ray structures.

4.6 Accuracy of IRECS when Predicting Flexible Side-Chain Conformations

The main ability of IRECS is to predict rotamer ensembles that represent the conformational space of all side chains. This ability of IRECS is much harder to evaluate than just comparing single predicted conformations to those found in X-ray structures. In this section an evaluation is attempted that displays the distribution of rotamers among all side chains of the protein and shows that the numbers of rotamers assigned to each side chain correspond to the assumed degree of flexibility of the respective side chains. A second evaluation compares the rotamers from rotamer ensembles generated with IRECS with conformational ensembles occasionally found in X-ray structures that are often but not always caused by the flexibility of side chains.

Table 4.3: Total numbers of compared rotamer ensembles of different sizes

Number of rotamers	Number of rotamers in crystal structure		
in IRECS model	1	2	≥ 3
1	79,790	1,947	13
2	30,761	2,051	37
≥ 3	10,113	825	22

4.6.1 Distribution of IRECS Rotamers

IRECS assigns different numbers of rotamers to each side chain and thus can weight their conformational flexibility. In Figure 4.6 the average number of rotamers per side chain is shown, resulting from 160 protein models built with IRECS on the native backbone and with $\rho_{rot} = 2$. Residues at the protein surface are assigned more rotamers than residues in the protein core. Amino acids that are known to be highly flexible like arginine, lysine, or glutamic acid are assigned a much higher conformational variability than amino acids known to be more rigid like tryptophan, tyrosine, cysteine, or histidine.

4.6.2 Accuracy of Predicted Rotamer Ensembles

The flexibility of side chains is hard to determine experimentally. With nuclear magnetic resonance (NMR) the conformational space of side chains can be detected, but this approach is limited to small proteins or peptide fragments of molecular weight below 40 kDa so far [20]. Fortunately, in a few cases crystallographers have attempted to fit multiple side-chain conformations to a measured electron density that does not allow for assigning a reasonable single conformation. These spread-out electron densities are assumed represent flexible side chains, but they can also arise from experimental artifacts like crystal packing or nearby disordered regions. Only if the multiple conformations of a side chain differed by at least 40° in either their χ_1 or their χ_2 dihedral angles, they were defined as multiple conformations for the purpose of this evaluation. The exclusion of NMR ensembles for this analysis makes sense, since these ensembles also contain multiple conformations of the backbone, which IRECS cannot handle. A representative subset of the PDB was constructed as a test set using the Protein Sequence Culling Server[5] (PISCES) [219, 220] of the Dunbrack Lab (see Appendix C.3). To ensure that multiple side-chain conformations do not originate from badly resolved crystallographic regions, only protein structures with a resolution better than 1.5 Å, an R and $\mathrm{R_{free}}$ below 0.3, and a pairwise sequence identity lower than 25% were chosen, as it was the case for the test set for single side-chain conformations (see Section 4.5). All structures without multiple side-chain conformations were discarded and 447 protein structures were kept. For each of these protein structures an IRECS model was built with a rotamer density of two.

In Table 4.3 the numbers of rotamers that are assigned by IRECS are shown together

[5] http://dunbrack.fccc.edu/PISCES.php

Table 4.4: χ_{1+2} accuracy of rotamer ensembles predicted with IRECS

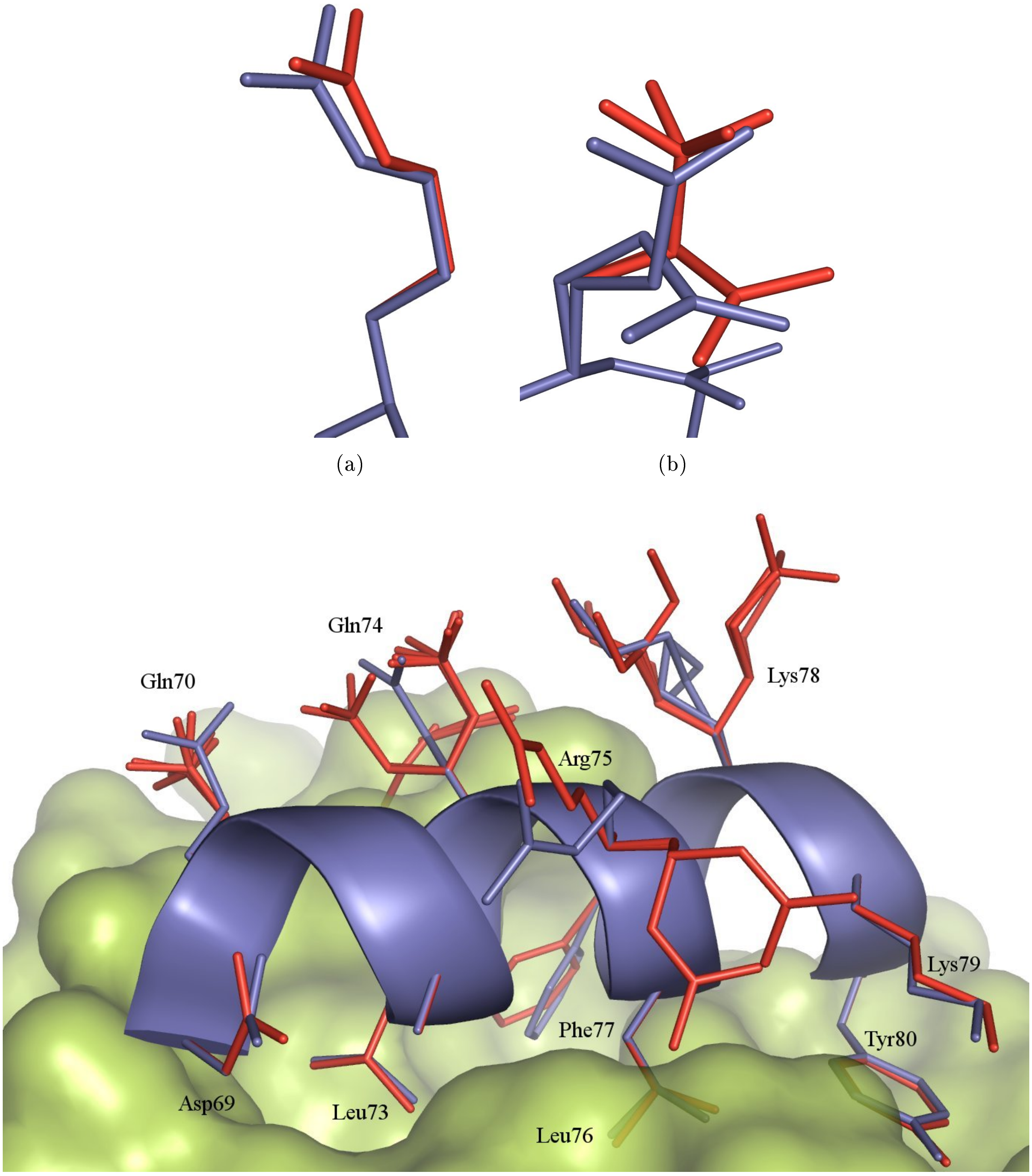
Number of rotamers in ensemble computed with IRECS	Number of matches between both ensembles / Number of rotamers in ensemble derived from crystal structure							
	1/1	1/2	2/2	$\geq 1/2$	1/3	2/3	3/3	$\geq 1/3$
1	88.7	81.2	3.9	85.1	92.3	0.0	0.0	92.3
2	87.1	38.7	54.7	93.4	21.6	75.7	0.0	97.3
≥ 3	85.0	28.7	64.5	93.2	4.5	9.1	81.8	95.4

This table lists the percentages of ensemble pairs that have certain sizes and a certain number of rotamers with similar $\chi_{1,2}$ dihedral angles (with a 40° cutoff).

with the number of rotamers that were found at this side chain in the original protein structure. It can be seen that IRECS generally assigns more rotamers to side chains that have multiple rotamers in their X-ray structures. The sizes of the ensembles assigned by IRECS do not match the sizes of the ensembles found in the protein structures perfectly, but this was expected as only a small fraction of flexible side chains actually is assigned multiple conformations in the crystal structures.

The structural agreement of multiple rotamers in the IRECS and crystal ensembles is used here to evaluate the ability of of IRECS to represent the conformational space of side chains. All rotamers in the IRECS ensembles were compared pairwise with the rotamers in the X-ray ensembles by the $\chi_{1,2}$ match criterion (see Section 4.4.2) and matches were counted as shown in Table 4.4. The table combines the sizes of the compared ensembles with the number of matching rotamers that could be found. A ratio (given in percentage) of ensemble pairs is shown that achieve the given number of matches among all ensemble pairs with the same size. For example, among all ensemble pairs for which the X-ray ensemble has two rotamers and the IRECS ensemble has two rotamers, 38.7% have one match and 54.7% have two matches. Whenever the X-ray ensembles hold two or three rotamers and IRECS decides to assign several rotamers to these side chains it manages to match one rotamer in 93.2% and 95.4% of these cases, respectively. This shows that IRECS is not only able to identify flexible side chains but also to construct rotamer ensembles that fairly well represent the conformational space of side chains. Whenever IRECS decides to assign only a single conformation to a side chain, it does so with high accuracy, since it hits the unique X-ray conformation in 88.8% of all cases, which is much higher than the accuracy of IRECS if it is restricted to using a single rotamer per side chain (71.6%, from Table 4.1). IRECS usually chooses to assign more rotamers to side chains that are harder to predict and can therefore maintain an overall good accuracy.

An example is shown in Figure 4.7. This figure shows a model with a rotamer density of two on the backbone of the human UDP-galactose 4-epimerase of crystal structure 1EK6 [205]. This crystal structure was chosen as an example because of its high quality (1.5 Å resolution, R-value 0.169, and R_{free} 0.198) and the presence of 16 side chains with alternative conformations. Figure 4.7 (a) shows a correctly predicted arginine; it points directly toward



(a)

(b)

(c)

Figure 4.7: IRECS model of human UDP-galactose 4-epimerase. The IRECS model (red) has rotamer ensembles with rotamer density two. The model was predicted on the backbone of the crystal structure of the human UDP-galactose 4-epimerase (blue, PDB code 1ek6, B chain) [205]. (a) Correct and unambiguous prediction of Arg61. This side chain points toward the core of the protein. (b) Prediction of Glu63 at the end of a beta-strand. Both ambiguous conformations of the crystal structure were predicted. (c) Helix on chain B. The helix has been cut out, and only the surface of the remaining crystal structure is drawn.

the protein core and therefore is highly restricted in its conformational space. Since IRECS assigns only one rotamer to this residue, Figure 4.7 (b) displays a glutamic acid at the protein surface with two alternative conformations in the crystal structure, which both were predicted by the IRECS algorithm. Figure 4.7 (b) shows the helix between residues 69 and 80 of the B chain. The conformational space of the residues on the helix surface ranges between four and nine rotamers, whereas all buried side chains have only one rotamer left. This figure is also intended to visualize some obstacles of side-chain prediction. For example, Gln74 has a χ_1 dihedral angle that is not represented by the rotamer library at all. IRECS selects the closest possible rotamers here, but all of the χ_1 differ by 60°. Arg75 has three rotamers assigned, with two of them being far from the native conformation and only one rotamer with a similar χ_1. Lys78 has two alternative conformations in the crystal structure, with occupancy of 0.5 each. IRECS assigns eight rotamers to this residue, matching one of the alternative conformations quite well and missing the second alternative conformation completely. All buried residues are predicted optimally, considering the granularity of the BBDep.

4.7 Application: Analyzing HCV Drug Resistance

An early application of IRECS was an analysis of possible rotamer states of a mutated protein: in a recent clinical trial mutations at 4 positions in the sequence of the hepatitis C virus (HCV) protease NS3-4A could be connected to drug resistance of HCV against the protease inhibitor VX-950 (Telepravir) [130, 174]. One of these mutations is Val36 to alanine, glycine, leucine or methionine [186]. A superposition of X-ray structures of HCV protease NS3-4A with STruster[6] [40] and structural comparison of the ligands with VX-950 (performed by Francisco Domingues) revealed close similarity between VX-950 and the ligand in the X-ray structure 1rtl [198], including a terminal cyclopropyl group. This group is buried in a subpocket of the protease binding site that is close to Val36. This gave rise to the starting hypothesis that mutations of Val36 affect this subpocket and triggers drug resistance. The following modeling procedure and analysis was carried out to validate this hypothesis [229].

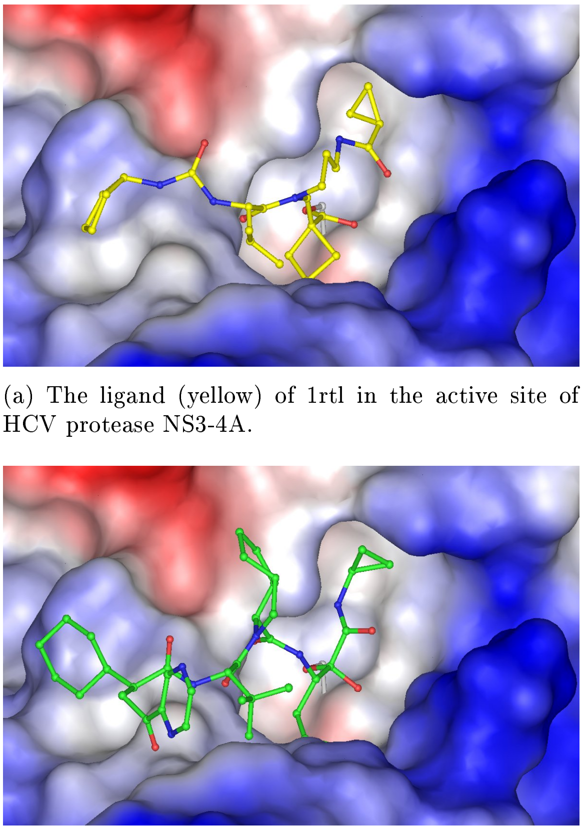
4.7.1 Docking VX-950 to HCV protease NS3-4A

To date no X-ray structure of the complex between VX-950 and HCV protease NS3-4A is publicly available, and so it was necessary to model this complex with protein-ligand docking. The structural similarity between VX-950 and the ligand in the X-ray structure 1rtl rendered this X-ray structure an optimal candidate for modeling the complex of VX-950 and the HCV protease NS3-4A. VX-950 was first drawn with MDL ISIS/Draw[7] and converted to 3D using energy minimization with the MMFF94 force field [68] as implemented in MOE[8]. VX-950 was docked with FlexX to the active site of the protease, using the ScreenScore [200] parameterization of FlexX. The cyclopropyl group of VX-950 was selected

[6] http://struster.bioinf.mpi-inf.mpg.de

[7] http://www.mdl.com/products/framework/isis_draw/

[8] http://www.chemcomp.com



(a) The ligand (yellow) of 1rtl in the active site of HCV protease NS3-4A.

(b) VX-950 docked to 1rtl

Figure 4.8: Visualizations of the active site of HCV protease NS3-4A. The surface of the protein was colored with the vacuum electrostatics function of PyMOL [35]. Charges are computed with the Amber99 force field [221] and projected on the protein surface, whereas colored patches (red=positive, blue=negative) denote polar regions and white patches apolar protein regions.

as the base fragment of FlexX to achieve a high sampling rate on this group. VX-950 was first docked into the active site without specifying a covalent bond. FlexX generates 9 different placements of VX-950, whereas the top ranking placement (see Figure 4.8(b)) shows a binding mode which is quite similar to that of the 1rtl ligand. The cyclopropyl group is placed towards the hydrophobic region in the protease ligand binding site, buried in the surface cavity and faces towards the aromatic ring of Phe43. The ketone oxygen of VX-950 is nearby the Ser139 side chain. Next, the covalent bond between this serine and the ketone oxygen was fixed and the structure of VX-950 was relaxed using a 100-step energy minimization with the MMFF94 force field. This two-step setup was chosen to ensure that the docking is not biased by any geometrical constraints that could occur when using a manually defined covalent bond during ligand placement.

4.7.2 Rotamer Analysis of Val36 Mutants

Rotamers of the mutated side chain of Val36 (mutations to alanine, glycine, leucine and methionine) were predicted with IRECS. Figure 4.9 illustrates possible side-chain conformations for the mutants of Val36. The conformational analysis of the mutants revealed that (i) there is only one atom in the gamma positions of the mutated side chains, (ii) all rotamers of the mutants are oriented uniformly towards the protein center and away from the ligand binding site and (iii) no carbon in the gamma position (C_γ) points towards the aromatic ring of Phe43, as it was the case in the wildtype structure. This carbon atom missing in the mutants interacts with Phe43 and restricts its flexibility [73]. Upon mutation, Phe43 can more easily change its conformation, which in turn affects the shape of the subpocket of the protease, to which the cyclopropyl group of VX-950 binds. This change can reduce the ability of VX-950 to bind to the protease, which can be an explanation for the lowered sensitivity of the respective viral variant to the drug observed during the clinical trial. This molecular explanation of drug resistance is supported by the observation that only such mutations were reported in the clinical trial that enable a single atom at the γ-position, and no mutations were reported that have two atoms at this position (e.g. the apolar isoleucine). However, the validity of the analysis undertaken here rests upon the accuracy of the modeled ligand conformation and the predicted conformations of Val36.

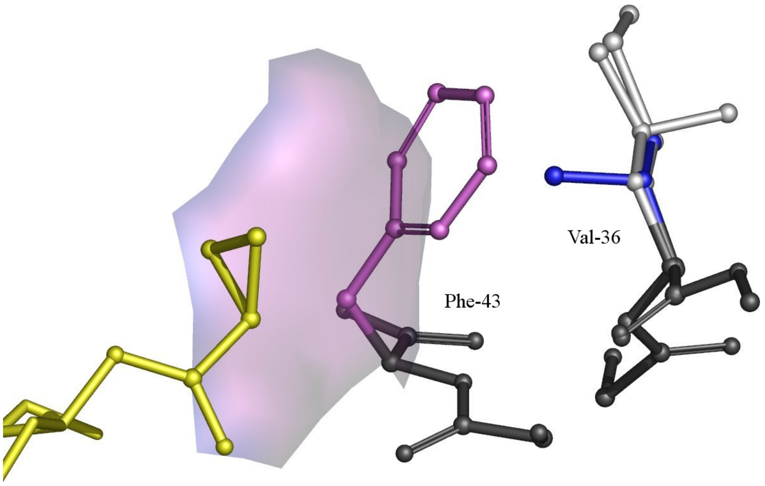


Figure 4.9: Mutants of Val36 in HCV Protease affecting the conformation of Phe43. This picture shows the main portion of the ligand of 1rtl (left, yellow), the amino acid Phe43 (middle, pink) and the amino acid Val36 (right, blue) and its mutants (right, white) alanine, glycine, leucine and methionine. The patch between the ligand and Phe43 illustrates that part of the solvent-accessible surface which depends on Phe43.

4.8 Discussion

The results presented here show that the accuracy of IRECS in predicting single side-chain conformations is in the same range as for well known, established methods, and therefore there is not much need for IRECS in this application area. However, the main benefits of IRECS are first its ability to detect side-chain flexibility and second to realistically sample the most important side-chain conformations. The ability of IRECS to predict the flexibility of side chains was shown in the previous section, but no direct comparison to other methods that also predict protein flexibility were performed. The main reason for this is that such methods are hard to compare since they concentrate either on predicting the vibrational displacement of atoms in X-ray structures [156, 235, 236] or conformational variability in multiple experimental structures [5, 40] or they predict protein flexibility through molecular dynamics simulations [188]. This usually includes estimating the flexibility of the backbone, which interferes with the computation of side-chain flexibility. However, the flexibility of the conformational ensembles that IRECS generates can be quantified in many ways, e.g. the number of generated rotamers, the variance of atom locations, or the average number of rotational degrees of freedom per side chain.

The heuristic nature of the IRECS algorithm renders it incapable of predicting rotamer ensembles that represent the full conformational space of side chains: in the case of two rotamers clashing with each other, one of the rotamers will be removed by IRECS in an early stage of the optimization. However, it is generally possible that both clashing rotamers are realistic representatives of the conformational space of their side chains, specifically if the one of the side chains adapts a non-clashing conformation. It is therefore not possible to generate certain alternative protein conformations from the IRECS ensembles that could also have quite low energy states.

An interesting property of the conformational ensembles generated with IRECS is that after the optimization is finished (e.g. with $\rho_{\mathrm{rot}} = 2$), the range of effective energies in the rotamer ensembles is low, which means that the contributions of the remaining rotamers to the effective energy is nearly the same. Another observation is that in most cases the interactions between the generated rotamers of two side chains and the interactions with the backbone have similar energetic contributions. As a consequence any selection of single rotamers from the rotamer ensembles generates rigid protein conformations with similar energy. The rotamer ensembles of IRECS can therefore serve as a pre-optimized protein configuration, from which rigid protein conformations can be generated. The computation of interactions between side chains can be neglected by the cost of a small amount of additional inaccuracy during energy computation. This property is later devised for docking into flexible proteins, see Chapter 6.

Chapter 5

Supporting Side-Chain Prediction with Structural Knowledge from Related Proteins

The basic concept of homology modeling is to use structural information from a homologous protein (the template) to model the structure of a target protein. After modeling of the backbone, it is common to model the conformations of all side chains or just the side chains of residues with changed amino acids. This is usually done with specialized tools like SCWRL [26], SCAP [233] or IRECS [72]. However, it is known that the accuracy of side-chain prediction tools strongly decreases if a modeled or a template backbone is used for the prediction instead of the backbone of the target protein [83, 210]. In this chapter, an algorithm called rotamer-lock algorithm is presented that allows for taking conformational knowledge from a template structure into account within the IRECS optimization procedure. In each optimization step of IRECS in which a rotamer should be removed (see Section 4.2.4) that is similar to that of the corresponding side chain in a template structure, it is decided whether IRECS should trust its own scoring function or take this similarity as a hint for rather protecting this rotamer from removal. The rotamer-lock algorithm uses two sets of classifiers that calculate for each dihedral angle (i) the probability that the part of rotamer defined by this dihedral angle is structurally similar to the respective part in the template protein and (ii) the probability that IRECS is able to predict the structure of this part correctly.

The first section of this chapter describes the rotamer-lock algorithm and its interaction with IRECS. The rotamer-lock algorithm requires the calculation of probabilities as noted above for which a set of decision trees is computed. A comprehensive data set for training and testing purposes is introduced thereafter, and a set of descriptive features that are required for the derivation of these decision trees is collected. Finally the performance of IRECS using the rotamer-lock algorithm is evaluated by 10-fold cross validation. The results are compared with the output of IRECS using a related algorithm, the conservation rule. Finally the advantages and disadvantages of the rotamer-lock algorithm and further methods for performance improvement of IRECS are discussed.

5.1 The Conservation Rule

The conservation rule is the most common method of incorporating side-chain information from a template. This rule directly transfers the conformation of a side chain from a template structure to a target model if the respective residue is conserved in the applied sequence alignment. The remaining side chains of mutated amino acids are then further optimized using some interaction-based scoring function, while the conserved side chains are being treated as having rigid conformations. This rule can be combined with nearly all side-chain optimization techniques that were developed so far. It first greatly simplifies the combinatorial problem of finding the optimal or near-optimal set of side-chain conformations, and second it usually enhances the performance of the overall side-chain prediction: in an evaluation of methods for model prediction, it was shown that applying the conservation rule clearly increases the accuracy of the predicted models with respect to other approaches that optimize all side chains [216]. The rule, however, has the drawback that it can only be applied to conserved residues, which decreases its usefulness if target and template protein exhibit little sequence similarity. It is also only of limited use for IRECS since it assigns just a single conformation to each conserved side chain and is not able to create ensembles of rotamers. The rotamer-lock algorithm was mainly developed to enable the inclusion of template information also in the generation of flexible protein models.

5.2 The Rotamer-Lock Algorithm

The idea of the rotamer-lock algorithm is to intervene in the IRECS optimization whenever a rotamer would be removed that is most similar to a template side chain within its rotamer ensemble. The rotamer-lock algorithm then checks if it is better to trust the decision of IRECS or rather to keep the side chain as it is in the template protein. These checks depend on the degree of similarity between the available rotamers in the ensemble and the template side chain: since the conformation of a side chain is determined by more than one dihedral angle for the majority of the standard amino acids, it is meaningful to determine structural similarity of side chains via the sequence of similar dihedral angles. This relation is hierarchical since the similarity of dihedral angles $\chi_1 - \chi_{i-1}$ is prerequisite for also concluding structural similarity of side chains by matching the dihedral angle χ_i. The rotamer-lock algorithm utilizes this hierarchy and tries to protect rotamers that are structurally similar with the template side chain at multiple levels.

The rotamer-lock algorithm is depicted in the flowchart of Figure 5.1. The reference dihedral angle χ_i of each protein side chain is determined first by the rotamer most similar to the respective template side chain and second by the presence of other similar rotamers in the current rotamer ensemble. Therefore the rotamer-lock algorithm protects not only one most similar rotamer from IRECS removal attempts (as the conservation rule would do) but larger rotamer sets of different size and similarity level. For example, arginine starts with 81 rotamers in IRECS from which only one has all dihedral angles $\chi_{1,2,3,4}$ matching with any given template side chain, but it has three rotamers with matching $\chi_{1,2,3}$ dihedral angles, nine rotamers with matching $\chi_{1,2}$ dihedral angles, and 27 rotamers with matching χ_1 dihedral

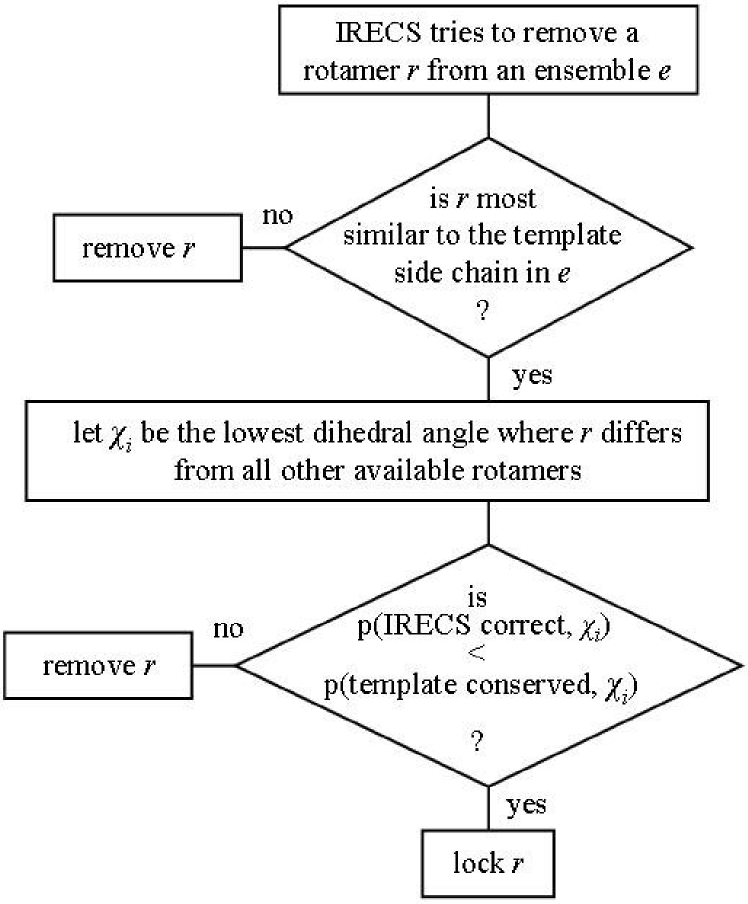

IRECS tries to remove a rotamer r from an ensemble e
is r most similar to the template side chain in e ?
no
remove r
yes
let χ_i be the lowest dihedral angle where r differs from all other available rotamers
is p(IRECS correct, χ_i) < p(template conserved, χ_i) ?
no
remove r
yes
lock r

Figure 5.1: Flowchart of the rotamer-lock algorithm

angle. The rotamer-lock algorithm attempts to protect these rotamers (i) whenever IRECS tries to remove one of them, (ii) if there is no active rotamer which is more similar, and (iii) if the probability of conservation of the side chain up to this dihedral angle is higher than the accuracy of IRECS for predicting all these dihedral angles correctly. The rotamer-lock algorithm then locks this rotamer. This has the effect that the search for a side-chain which should loose a rotamer (see Section 4.2.4) is repeated, this time ignoring the side chain with the locked rotamer (and all other side chains with locked rotamers) during the calculation of effective energy ranges (see Equation 4.9). The lock is maintained until either the evaluated probabilities or the reference dihedral angle χ_i changes. The probabilities can change during the IRECS optimization since they depend on features which assess the agreement between the current stage of the IRECS optimization algorithm and the respective side chain in the template protein. The reference dihedral angle changes when other rotamers are removed from the ensemble so that the current reference dihedral angle is not longer the lowest in sequence by which the current most similar rotamer differs from the other rotamers in its rotamer ensemble.

An example of this situation is depicted in Figure 5.2. This figure shows an ensemble of

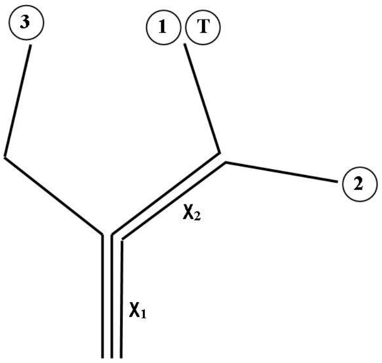


Figure 5.2: Example for an application of the rotamer-lock algorithm. The sketch shows a side chain with two dihedral angles and three rotamers remaining in the IRECS ensemble. Rotamer 1 has the same conformation as the respective side chain (T) in a template structure.

three rotamers, with rotamer 1 being similar to the template rotamer T for dihedral angles χ_1 and χ_2 and rotamer 2 having a similar χ_1 dihedral angle. At first, the reference dihedral angle is χ_2, since rotamers 1 and 2 differ in this dihedral angle. If after some time IRECS removes rotamer 2, the reference dihedral angle becomes χ_1 since this is now the lowest (in sequence) dihedral angle in which rotamer 1 differs from rotamer 3.

5.3 Data Set Assembly: Selection of Homologous Protein Pairs

The protein chains for deriving the required decision trees and performing the final evaluation were selected using a multi-step filtering procedure. The aim of the following filtering procedure is to get a representative set of pairs of homologous protein chains with a large number of aligned side chain pairs. Structures of low quality or peptide fragments should be excluded. The data set will be later used to train classifiers that predict the chance of two side chains being structurally similar and the chance of a IRECS prediction being correct.

First, the pairwise sequence identity of protein chains was calculated using the structural alignments of the DALI Fold database [79, 80]. The protein chains of the PDB were then clustered by their pairwise sequence identity, so that each chain has a pairwise sequence identity of less than 30% with all protein chains from other clusters. Each protein chain in the clusters must then have a resolution better than or equal to 2.0 Å, a length of at least 80 residues and must be resolved by X-ray crystallography or it will be deleted from the clusters. From each of the resulting clusters a pair of similar protein chains was extracted. Six groups of protein chain pairs were defined according by the respective pairwise sequence identity, 30%-39%, 40%-49%, 50%-59%, 60%-69%, 70%-79% and 80%-90% (percentages rounded to integers, see Appendix C.4.1-C.4.6 for a list of PDB ids for each group). Exactly one pair of protein chains was added to one of the groups from each protein clusters, if it fulfilled the following conditions:

1. there is an alignment for the chains in the DALI Fold database,

2. the backbones of both structures have an RMSD < 2.5 Å after superposition,
3. the sequence identity is at most 90%,
4. less than 20% of all residues in each chain are aligned against gaps and
5. both proteins have the same number of chains.

If there are candidate pairs for more than one group in a cluster, pairs were chosen so that the sizes of the six groups are balanced in size as possible. This procedure generated a total of 584 protein chain pairs. For each protein chain an IRECS model was created with single side chains, taking the backbone conformation from each respective partner chain as template. This procedure allowed each of the proteins to take on the role of the template and the role of the target protein, respectively. A number of 228,925 side-chain triplets were obtained with this procedure, which consists of a pair of corresponding side chains from related proteins with one marked as template and the other as target and a third side chain that is part of an IRECS model of the target protein. Table 5.1 summarizes the most important features of the data set.

5.4 Generation of Decision Trees

Two features of each side chain in the target protein must be predicted: (i) if the side chain has the same conformation in template and target (or short: *template conservation*), and (ii) if the conformation of a side chain is predicted correctly by IRECS (or short: *IRECS correct*). This results in two binary classification problems. Each of problem must be solved for each of the four levels of structural similarity of side chains (χ_1, $\chi_{1,2}$, $\chi_{1,2,3}$, $\chi_{1,2,3,4}$) which results in eight different classification tasks.

The software package WEKA[1] (Waikato Environment for Knowledge Analysis) [230] offers a large variety of classification and regression methods. Among these, 10 different classification algorithms that are either tree- or rule-based were evaluated with respect to their ability to predict if the χ_1 angle is conserved between corresponding side chains in

[1] http://www.cs.waikato.ac.nz/ml/weka/

Table 5.1: Total numbers and percentages for all side chains in the data set

	Percentage	Total number
Residues in dataset	100.0	228,925
Residues conserved	53.9	123,320
χ_1 conserved*	65.1	149,023
IRECS χ_1 correct*	71.6	163,954

* the respective residues of both the template structure and the query structure have the same χ_1 -angle slot (*gauge*$^+$, *gauge*$^-$ or *trans*) as defined for the BBDep (see Section 2.3.2).

Table 5.2: AUC for ten different classification algorithms

Algorithm	Nodes/Rules	AUC
J48 [164]	2,263	0.823
Naive Bayes Tree	615	0.812
Alternating Tree [54]	31	0.800
Decision Table [113]	15,431	0.768
Conjunctive Rule	2	0.703
Decision Stump	3	0.702
RIpple-DOwn Rule Learner [57]	93	0.600
OneR [81]	2,012	0.538
REPTree	443	0.500
ZeroR	1	0.500

Classifiers without a reference are explained with appropriate detail in the WEKA documentation [230].

a pair of homologous protein structures, as it is shown in Table 5.2. The J4.8 algorithm (successor of C4.5 [164] and ID3 [165], implemented in Java) was chosen due to its superior performance on this test. Also, an algorithm based on decision trees was preferred, because it enables easy interpretation of the predictions and is known to have a good performance if both categorical and numerical features are provided [166]. The optimal settings for learning trees for both problems were found by manual search through the parameter space of the algorithm:

1. tree nodes may have more than two child nodes,

2. no pruning of the tree and

3. a node may only be split further if more than 100 instances of the training set is assigned to it.

Given a query instance, the probability that this instance is a member in a certain class is computed by following a path through the decision tree that is determined by the feature values of the query instance. The leaf of this path holds a subset of the training data set, and the ration of the training instances belonging to the class to be predicted divided by all instances assigned to this leaf yields the estimate of the probability for the query instance to belong to the respective class.

5.4.1 Selection of Features for Classification

Seventeen features were selected, based first on the sequence and structure of both the template and target structure and second on the models generated with IRECS. All selected features and their possible values are listed in Table 5.3. The features are grouped by the source of information that is required to determine the value of the respective feature.

Table 5.3: Overview of all features used in the classification

Origin/Type	Feature	Range
template backbone	local number of C_β atoms	0- 20
template backbone	local sequence identity	0%-100%
template backbone	global sequence identity	0%-100%
template backbone	backbone Φ and Ψ	-180°-180°
template backbone	local secondary structure	loop, helix,strand
template sequence	amino acid name	Ala-Val
template sequence	amino acid type	al, uc, ch, ar, su
template sequence	number of dihedrals	0,1,2,3-4
target sequence	amino acid name	Ala-Val
target sequence	amino acid type	al, uc, ch, ar, su
target sequence	number of dihedrals	1,2,3,4
template side chains	χ slots	0,1,2,3
IRECS side chains	χ slots of most probable rotamer	0,1,2,3
indicator	mutation occurred	0,1
indicator	size changed	0,1
indicator	chemical type changed	0,1
indicator	IRECS and template vote for same χ	0,1

al = aliphatic, uc = uncharged, ch = charged, ar = aromatic, su = sulfur containing

These are the template and target sequence, the template backbone and side chains, the side chains of the IRECS model and some additional indicators. These features are used to characterize (i) the kind of mutation (if any) that occurs at a certain residue, (ii) the structural environment of this residue and its potential difference in the template and the target and (iii) the specific conformation of the target side chain both assigned by IRECS and observed in the template protein. χ-slots were used to fragment the torsional space of all dihedral angles of the side chains into discrete parts: the slot is set to 1 for χ in the interval $[0°, +120°[$, 2 for $[+120°, -120°[$, 3 for $[-120°, 0°]$ and 0 if the side chain does not have such a dihedral angle. The local sequence identity and the local number of C_β atoms were computed in each spatial neighborhood of a residue. This neighborhood comprises all residues that have their C_β atoms located within 10.0 Å of the relevant C_β atom of the residue (C_α is taken instead for Glycine residues). As Jones [95] states, the relative solvent accessibility (see Section 4.4.3) of a residue in a folded protein correlates highly (correlation coefficient 0.85) with the number of C_β atoms in such a sphere. As it was shown in the evaluation of IRECS, this property is quite determining for the accuracy of side-chain prediction, although it was also observed that this feature is quite uncorrelated with accuracy if a template backbone is used [192]. The local secondary structure was determined using the Φ and Ψ dihedral angles of the particular residue and localizing the corresponding secondary structure element on the Ramachandran map [167, 178, 22]. A number of additional features were introduced as indicators of certain events and are derived from other features (e.g. some feature exceeding

Table 5.4: Features ranked by their information gain ratio

	IRECS $\chi_{1,2}$ correct		Template $\chi_{1,2}$ conserved	
Feature	Rank	Gain ratio	Rank	Gain ratio
local number of C_β atoms	11	0.0088	13	0.0100
local sequence identity	13	0.0038	12	0.0135
global sequence identity	17	0.0010	14	0.0054
template backbone Φ	14	0.0034	15	0.0026
template backbone Ψ	16	0.0012	17	0.0008
local secondary structure	18	0.0005	18	0.0002
template amino acid name	5	0.0227	8	0.0523
template amino acid type	7	0.0154	9	0.0307
template number of dihedrals	12	0.0069	6	0.0777
target amino acid name	3	0.0349	10	0.0285
target amino acid type	6	0.0163	11	0.0212
target number of dihedrals	15	0.0029	16	0.0023
template $\chi_{1,2}$	4	0.0238	4	0.0982
IRECS $\chi_{1,2}$	2	0.0666	7	0.0565
mutation occurred	9	0.0134	2	0.1234
size changed	8	0.0139	3	0.1082
chemical type changed	10	0.0118	5	0.0904
template $\chi_{1,2}$ = IRECS $\chi_{1,2}$	1	0.1432	1	0.2270

a threshold) and thus have a high correlation with them. These indicators can prevent the J4.8 algorithm from fragmenting the data set too early. One example: the indicator 'mutation occurred' splits the data set into halves, whereas the same information can also be derived from the features 'amino acid name' of target and template. However, using the features 'amino acid names' of target and template in combination splits the data set into 20 times 18 = 360 fragments (alanine and glycine are not considered as target amino acid types during this analysis).

5.4.2 Benchmarking Single Feature Performance

The usefulness of these features was measured by the information gain ratio of each feature. This ratio measures the ability of a certain feature to fragment the training set into groups that contain clear majorities of instances that are member in one of the target classes (gain of information), but favors features that fragment the training set into a small set of groups. The gain ratio serves as the split criterion for the J4.8 decision tree algorithm [164, 165]. A precise formulation is given in an early publication of Quinlan [164]. The gain ratio takes into account (i) the gain of information if a certain feature is used to partition a given set of data instances and (ii) the (undesirable) fragmentation of the data set through the application of this feature for partitioning. Table 5.4 lists the features with their gain ratios and their rank when the features are be sorted by decreasing gain ratio. The ranking was

Table 5.5: Performance of decision trees

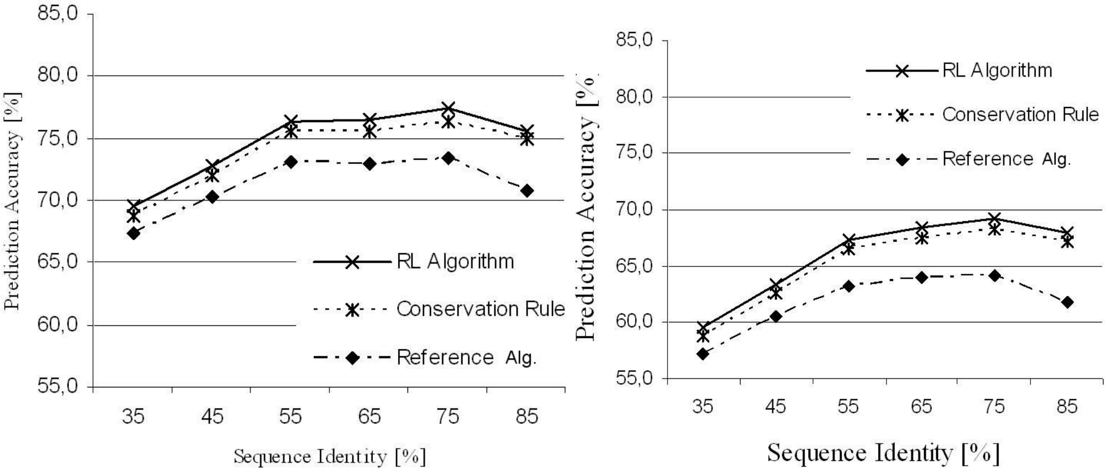
Decision Tree	Training Set AUC	X-validation AUC
IRECS χ_1 correct	0.7955	0.7854
IRECS $\chi_{1,2}$ correct	0.8288	0.8205
IRECS $\chi_{1,2,3}$ correct	0.7290	0.7189
IRECS $\chi_{1,2,3,4}$ correct	0.7357	0.7267
conserved template χ_1	0.8305	0.8243
conserved template $\chi_{1,2}$	0.8926	0.8878
conserved template $\chi_{1,2,3}$	0.8426	0.8365
conserved template $\chi_{1,2,3,4}$	0.8570	0.8499

Performance of various decision trees for prediction of congruency between side-chain dihedrals in protein structures. The trees with the prefix 'IRECS' predict the correspondence between the side chains of crystal structures and models generated with IRECS for the respective dihedral angles. Those trees with the prefix 'Template' predict the correspondence between the side chains of crystal structures with their template structures for the respective dihedral angles.

performed for the classification tasks '*IRECS correct*, $\chi_{1,2}$' and '*template conserved*, $\chi_{1,2}$'. The ranking shows that the importance of the features is quite similar for both classification tasks, both rankings have a rank correlation coefficient of 0.718 (see Equation 3.7). Global attributes that are meant to describe the overall environment (global and local sequence identity, local number of C_β atoms, conformation of the backbone) are less important than those features that characterize the conformation of the template side chain and the current state of the IRECS optimization (template $\chi_{1,2}$ = IRECS $\chi_{1,2}$, indicator that the most probable rotamer of IRECS has the same $\chi_{1,2}$ dihedral angles than the template side chain). This similar ranking points to the later difficulty of discriminating both classes from each other.

5.5 Performance of Decision Trees

The predictive power of the eight decision trees was measured by their performance on the full training set and by 10-fold cross-validation on the training set. The comparison of both evaluations allows for estimating the training error. Each decision made by a tree results in a probability for one of the classes. By sorting the probabilities for all queries and using multiple probability cutoffs for class assignment, Receiver-Operator Characteristics (ROC) curves [74] for all trees were created. Table 5.5 summarizes the properties of the trees and their performance, measured by the area under the curve (AUC) for evaluation on the full training set and the partial training set as defined by the cross-validation. The AUC ranges from 0.719 to 0.889 in the cross validation and is highest for the trees predicting the template $\chi_{1,2}$ angle. The trees that were derived for predicting the performance of IRECS have lower performance than the trees that were derived for the prediction of side-chain conservation.



(a) χ_1 match criterion

(b) $\chi_{1,2}$ match criterion

Figure 5.3: Accuracy of IRECS using the rotamer-lock algorithm The accuracy of IRECS is shown for different target/template pairs grouped by their sequence identity when (i) using no additional knowledge from a template structure, (ii) using the conservation rule and (iii) the rotamer-lock algorithm to protect rotamers that are similar to the correpsonding template side chain from IRECS removal.

In general, the results of the cross validation are just slightly worse than the results of the validation on the full training data set, which suggests that the setup of features and decision trees successfully prevented overtraining. This can also result from the low impact that an instance of the data set has on the final structure of the trees, since the split criterion prevents splitting a node further into subtrees if it represents fewer than 100 instances. The decreasing accuracy for the $\chi_{1,2,3}$ and $\chi_{1,2,3,4}$ dihedral angles may primarily result from the natural flexibility of such long side chains.

5.6 Accuracy of IRECS with the Rotamer-Lock Algorithm

The performance of two alternative methods for incorporating side-chain information from the template into the IRECS modeling procedure, the conservation rule and the rotamer-lock algorithm, has been evaluated on the data set. This data set is also used in training the decision trees, and therefore the test is performed using 10-fold cross validation like in Section 5.5. To estimate the effect that both methods have on the prediction accuracy, these results are compared to the performance of the original IRECS algorithm without template information (called reference in the following). The results are depicted in Figure 5.3, which shows the percentage of matching χ_1 and $\chi_{1,2}$ dihedral angles of predicted rigid side chains with side-chain conformations of rigid X-ray structures. Since the side-chain information from the template becomes more important with increasing structural similarity between template and target protein, the performance is evaluated for the six different groups of homology. As expected the accuracy of all three IRECS implementations increases as the sequence similarity between template and target increases. Although higher similarity

Table 5.6: Prediction accuracy of IRECS for different amino acids

	Applied algorithm and target rotamer density				
Amino	Reference algorithm		Conservation rule	RL algorithm	
acid	$\rho_{\text{rot}} = 1$	$\rho_{\text{rot}} = 2$	$\rho_{\text{rot}} = 1$	$\rho_{\text{rot}} = 1$	$\rho_{\text{rot}} = 2$
Arg	46.1	59.8	50.3	**52.1**	**61.0**
Asn	53.3	71.2	**59.7**	58.2	**71.9**
Asp	59.1	73.7	63.4	**64.1**	**74.9**
Cys	72.8	85.7	80.3	**80.7**	**88.8**
Gln	44.1	62.4	48.7	**49.2**	**63.7**
Glu	40.8	58.0	44.0	**45.3**	**60.0**
His	63.5	76.9	**71.6**	70.0	**80.4**
Ile	62.7	71.1	64.4	**65.5**	**72.7**
Leu	74.5	81.3	73.4	**75.8**	**81.7**
Lys	44.0	60.5	44.4	**46.5**	**60.8**
Met	50.6	68.4	55.5	**56.0**	**69.9**
Phe	76.0	81.7	**80.4**	**80.4**	**84.3**
Pro	64.5	81.5	66.9	**67.4**	**83.3**
Ser	53.3	80.3	58.3	**60.0**	**82.2**
Thr	71.0	78.7	72.0	**72.7**	**79.6**
Trp	58.2	64.8	**79.1**	76.4	78.8
Tyr	78.9	83.4	**83.1**	83.0	**86.0**
Val	**97.2**	**97.4**	96.2	**97.2**	**97.4**
All	62.9	75.1	65.9	**66.8**	**76.7**

The $\chi_{1,2}$ accuracy is given for each amino acid and all applied algorithms for building rigid and flexible protein models. Cells with highest accuracy compared to other prediction runs with same rotamer density are highlighted.

between target and template structure generally eases side-chain prediction, the accuracy of all three IRECS version drops in the group of highest homology (80-90%). Such a drop was also reported in a similar evaluation made by Wallner and Elofsson [216]. This phenomenon can be partially explained by the fact that in some cases sequence identity is an insufficient measure for the difficulty of a prediction. Also this part of the data set has the fewest protein structures compared to the other parts, which promotes outliers. Apart from this effect, the figures support the general usefulness of the conservation rule and the rotamer-lock algorithm. The rotamer-lock algorithm always outperforms the conservation rule and allows IRECS to improve its average prediction accuracy from 72.7% to 75.9% for the χ_1 dihedral angle and from 62.9% to 66.8% for the $\chi_{1,2}$ dihedral angles on this data set.

The accuracy of different IRECS implementations (with and without the rotamer-lock algorithm) has also been evaluated with respect to their ability to generate accurate protein models with multiple side-chain conformations. Since the number of resolved side chains with multiple conformations in crystal structures is much lower in this data set than in the previously presented test set (see Section 4.6.2), the IRECS ensembles are compared to single conformations of side chains in crystal structures. Instead of a one-by-one matching of side chains in the model and the reference structure respectively, a side-chain conformation is considered matched if one of the rotamers in the ensemble matches it, as introduced in Section 4.4.2. This is a drastic simplification of the prediction problem for single conformations, therefore only predictions with the same rotamer density can be compared with each other. The average accuracy for each amino acid type is shown in Table 5.6. For models with a rotamer density of one, the conservation rule leads to a better average accuracy than the rotamer-lock algorithm in six cases, nearly equal accuracy for three amino acid types and worse accuracy for nine amino acid types. The average accuracy among all amino acid type is also highest for the rotamer-lock algorithm. For models with a rotamer density of two, the rotamer-lock algorithm always increased the accuracy of IRECS except for valine, by a maximal value of 14% for tryptophan.

5.7 Potential and Limitations of the Approach

The rotamer-lock algorithm presented here just slightly increased the accuracy of IRECS. Two issues appear to be responsible for this low performance of the rotamer-lock algorithm. First, the applied data set was not designed first-hand for this kind of analysis. Another training set could potentially improve the derived decision trees. Such a data set could contain more recent X-ray structures (the data set of Holm et al. [80] is already ten years old), pair alignments would be based on more recent approaches and side chains would be excluded if the temperature factor of its atoms indicate that no reliable rotamer assignment could be performed during fitting the electron density. A second issue is that the original problem – deciding whether to trust the scoring function or assume conformational conservation between template and target side chain – was artificially split into two separate prediction problems. As both problems are determined by quite similar sets of features (see Figure 5.4), this setup prevented the most discriminatory features from being identified by the J4.8 algorithm.

Nevertheless, we could show the general usefulness of the idea of repeatedly predicting the accuracy of an optimization algorithm while it is running and of improving the optimization procedure with the use of additional knowledge. This idea can also be beneficial in other applications. As we do not compute conformations but only protect certain rotamers from early removal, one can combine this approach with other side-chain prediction tools that utilize rotamer-reduction like the R3 algorithm [234]. It can be expected that the overall performance of the rotamer-lock algorithm can be increased further (i) with an updated data set, (ii) a suitable feature representing residue-specific structural deviations between target and template backbone, (iii) the extension to multiple aligned template structures and template rotamers, (iv) by derivation of a single classifier that directly decides which of the two classes has higher probability and/or (v) the training and optimization of other types of classifiers like support-vector machines or neural networks.

Chapter 6

Docking with Flexible Side Chains

This chapter addresses the final question asked in the introduction (see Section 1.1): 'how can side-chain flexibility be included during docking so that a high amount of protein binders can be identified during virtual screening?' A modeling and docking pipeline is described in this chapter that is able to deal either with proteins that exhibit induced-fit effects upon ligand binding or with proteins for which only models with inaccurate side-chain conformations are available. With the help of this pipeline multiple redocking and screening experiments are performed. An evaluation is made that shows that the side-chain conformations generated with IRECS are sufficiently accurate for docking to succeed. The results allow also for comparing different parameterizations of IRECS and FlexE to handle protein flexibility. These are three primary aspects of our study:

1. which rotamer density to choose for the models,
2. whether or not IRECS should use ligand information to pre-optimize side-chain ensembles for FlexE and
3. which scoring function, ROTA or F-Score, is better suited for docking into flexible protein models.

A special focus lies on the first aspect, since the value of rotamer density of the IRECS models has the largest influence on the later handling of protein flexibility by FlexE. This pipeline is shown in Figure 6.1. Starting from an X-ray structure, IRECS is used to repredict the side-chain conformations. Depending on the protein flexibility or the difficulty of assigning the correct side-chain conformations, IRECS builds protein models with varying side-chain flexibility. The docking is carried out with the docking program FlexE in case of flexible protein models and with FlexX in case of rigid protein models. FlexX and FlexE then use either the usual scoring function of FlexX (F-Score) [172] or ROTA for guiding the docking process and predict the final binding affinity of the ligand to the protein.

An analysis is carried out which should give an answer to the final question asked in the introduction. It is shown that the proposed docking pipeline is able to handle side-chain flexibility efficiently. The screening experiments facilitate determining the appropriate amount of side-chain flexibility for obtaining high enrichment factors in general and also for specific proteins.

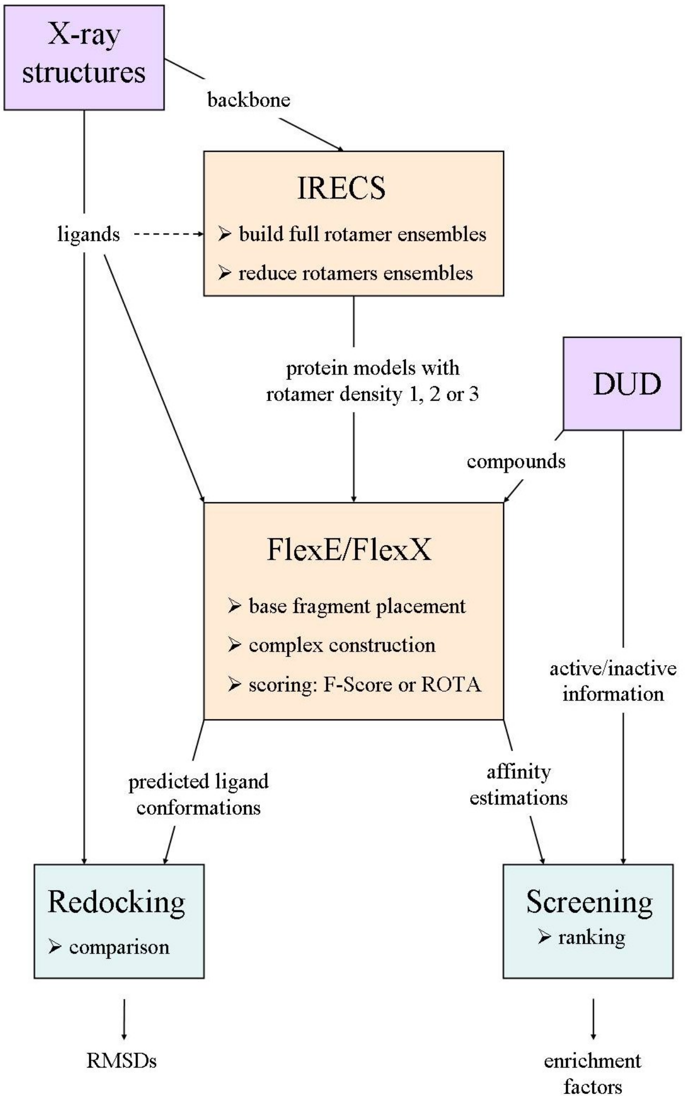

X-ray structures
backbone
IRECS
➢ build full rotamer ensembles
➢ reduce rotamers ensembles
ligands
protein models with rotamer density 1, 2 or 3
DUD
compounds
FlexE/FlexX
➢ base fragment placement
➢ complex construction
➢ scoring: F-Score or ROTA
active/inactive information
predicted ligand conformations
affinity estimations
Redocking
➢ comparison
Screening
➢ ranking
RMSDs
enrichment factors

Figure 6.1: Data flow of the modeling and docking pipeline

The first section of this chapter gives a detailed description of the modeling and docking pipeline. The second section then describes the setup and results of the redocking and screening experiments. A comparison of the method proposed here with other methods is presented in the last section. A discussion of the docking pipeline is given in the final discussion in the next chapter, since the pipeline comprises the material and methods from the previous chapters and should be discussed in that global context.

6.1 The Modeling and Docking Pipeline

Figure 6.1 shows the modeling and docking pipeline and how experimental data are used for running both redocking and virtual screening experiments. Since ROTA, IRECS and FlexE were previously benchmarked as stand-alone programs, the evaluation here focuses on how well both programs can perform in combination.

6.1.1 Step 1: Extraction of Experimental Data

The Database of Useful Decoys (DUD) [85] comprises forty proteins that are well-known drug targets. For each of these proteins Huang and coworkers stored an X-ray structure, a set of active compounds (called *ligands*) and a set of putative inactive compounds (called *decoys*) in DUD. Table 6.1 lists all proteins of the DUD together with the abbreviations that are used from now on, the PDB code and the resolution of an X-ray structure and the numbers of ligands and decoys that are contained for the respective protein in the DUD. For VEGFR2 the X-ray structure 1y6a was used instead of 1vr2 since in 1y6a the protein was crystallized in complex with a ligand, whereas the protein in 1vr2 is in the apo form. Huang and coworkers extracted decoys from the drug-like subset of the ZINC database [90] such that for each ligand there are about 35 decoys which are similar in physical properties (e.g. molecular weight, number of hydrogen bond acceptors or logP) but are topologically different. This setup ensures that this is hard to identify active compounds just by such simple physical properties.

6.1.2 Step 2: Building Protein Models with IRECS

IRECS is used to build multiple models of all proteins that are used in the later evaluations. For each protein six models were built with IRECS: a first subset of three models was built with rotamer densities of one, two and three, respectively, based only on the backbone conformation of the respective protein. IRECS models with higher rotamer densities (five and seven) were also build in preliminary tests. Since their docking results did not show significant differences from those docking results achieved when using IRECS models with a rotamer density of three, these models were not used in further evaluations. A second subset of three models was created again with these three rotamer densities and using the backbone conformation, but now the conformation of the native ligand in the pocket was kept rigid and was regarded as part of the protein during the optimization: IRECS computes interactions between each rotamer x_i and the ligand and adds this interaction value as a new term $U_{\text{inter}}(x_i, \text{ligand})$ to the effective rotamer energy (see Equation 4.7) of the respective rotamer.

Table 6.2 lists the average accuracy of the protein models from the six different IRECS runs in the active site. All side chains of the X-ray structures in the active sites were compared to their counterparts in the IRECS models by χ-matching (see Section 4.5.1) and RMSD, respectively. The active sites were defined with the help of the ligand conformations of the X-ray structures so that any residue of the protein that has an atom within 6.5 Å distance to a ligand atom is considered as belonging to the active site. The average χ-match and RMSD values of the individual proteins were averaged over all generated models of a single IRECS parameterization. It can be seen that the average side-chain RMSDs of the IRECS models decrease and the χ-match values increase if ligand information is used during the optimization. The side-chain accuracy also trivially increases with increasing rotamer density, since more rotamers become available for matching in these models. The IRECS

Table 6.1: Properties of the protein targets of the DUD [85]

	Abbrev-	PDB	Reso-	No. of	No. of
Protein	iation	code	lution [Å]	ligands	decoys
Angiotensin-converting enzyme	ACE	1o86	2.0	49	1728
Acetylcholine esterase	AChE	1eve	2.5	105	3732
Adenosine deaminase	ADA	1ndw	2.0	23	822
Aldose reductase	ALR2	1ah3	2.3	26	920
AmpC β-lactamase	AmpC	1xgj	2.0	21	734
Androgen receptor	AR	1xq2	1.9	74	2630
Cyclin dependent kinase 2	CDK2	1ckp	2.1	50	1780
Catechol O-methyltransferase	COMT	1h1d	2.0	12	430
Cyclooxygenase 1	COX-1	1p4g	2.1	25	850
Cyclooxygenase 2	COX-2	1cx2	3.0	349	12491
Dihydrofolate reductase	DHFR	3dfr	1.7	201	7150
Epidermal growth factor receptor kinase	EGFr	1m17	2.6	416	14914
Estrogen receptor agonist	$\mathrm{ER}_{\mathrm{agonist}}$	1l2i	1.9	67	2361
Estrogen receptor antagonist	$\mathrm{ER}_{\mathrm{antagonist}}$	3ert	1.9	39	1399
Fibroblast growth factor receptor kinase	FGFr1	1agw	2.4	118	4216
Factor Xa	FXa	1f0r	2.7	142	5102
Glycinamide ribonucleotide transformylase	GART	1c2t	2.1	21	753
Glycogen phosphorylase β	GPB	1a8i	1.8	52	850
Glutocorticoid receptor	GR	1m2z	2.5	78	2804
HIV protease	HIVPR	1hpx	2.0	53	1888
HIV reverse transcriptase	HIVRT	1rt1	2.6	40	1439
Hydroxymethylglutaryl-CoA reductase	HMGR	1hw8	2.1	35	1242
Human heat shock protein 90 kinase	HSP90	1uy6	1.9	24	861
Enoyl ACP reductase	InhA	1p44	2.7	85	3043
Mineralcorticoid receptor	MR	2aa2	1.9	15	535
Neuraminidase	NA	1a4g	2.2	49	1745
P38 mitogen activated protein kinase	P38 MAP	1kv2	2.8	234	8399
Poly(ADP-ribose) polymerase	PARP	1efy	2.2	33	1178
Phosphodiesterase V	PDE5	1xp0	1.8	51	1810
Platelet derived growth factor receptor β	PDGFrb	model	n.a.	157	5625
Purine nucleoside phosphorylase	PNP	1b80	1.5	25	884
Peroxisome proliferator activated receptor γ	PPARg	1fm9	2.1	81	2910
Progesterone receptor	PR	1sr7	1.9	27	967
Retinoic X receptor α	RXRa	1mvc	1.9	20	708
S-adenosyl-homocysteine hydrolase	SAHH	1a7a	2.8	33	1159
Tyrosine kinase C-SRC	SRC	2src	1.5	162	5801
Thrombin	thrombin	1ba8	1.8	65	2294
Thymidine kinase	TK	1kim	2.1	22	785
Trypsin	trypsin	1bju	1.8	43	1545
Vascular endothelial growth factor receptor	VEGFr2	1y6a	2.4	74	2647

Table 6.2: Accuracy of the modeled side chains in the active sites of the DUD targets

Rotamer density	Ligand info	Side-chain RMSD [Å]	Standard deviation [Å]	Accuracy of dihedral angles χ_1[%]	χ_{1+2}[%]
1	no	1.15	0.29	83.2	79.8
1	yes	1.04	0.25	85.3	82.3
2	no	0.88	0.27	89.8	81.2
2	yes	0.85	0.23	91.0	81.0
3	no	0.77	0.25	92.4	86.4
3	yes	0.75	0.21	93.5	86.2

RMSD values and χ-match percentages were measured in the active site of the respective models. The active sites covered about 25 residues on average (with a standard deviation of 5.5 residues) within all six sets of IRECS models.

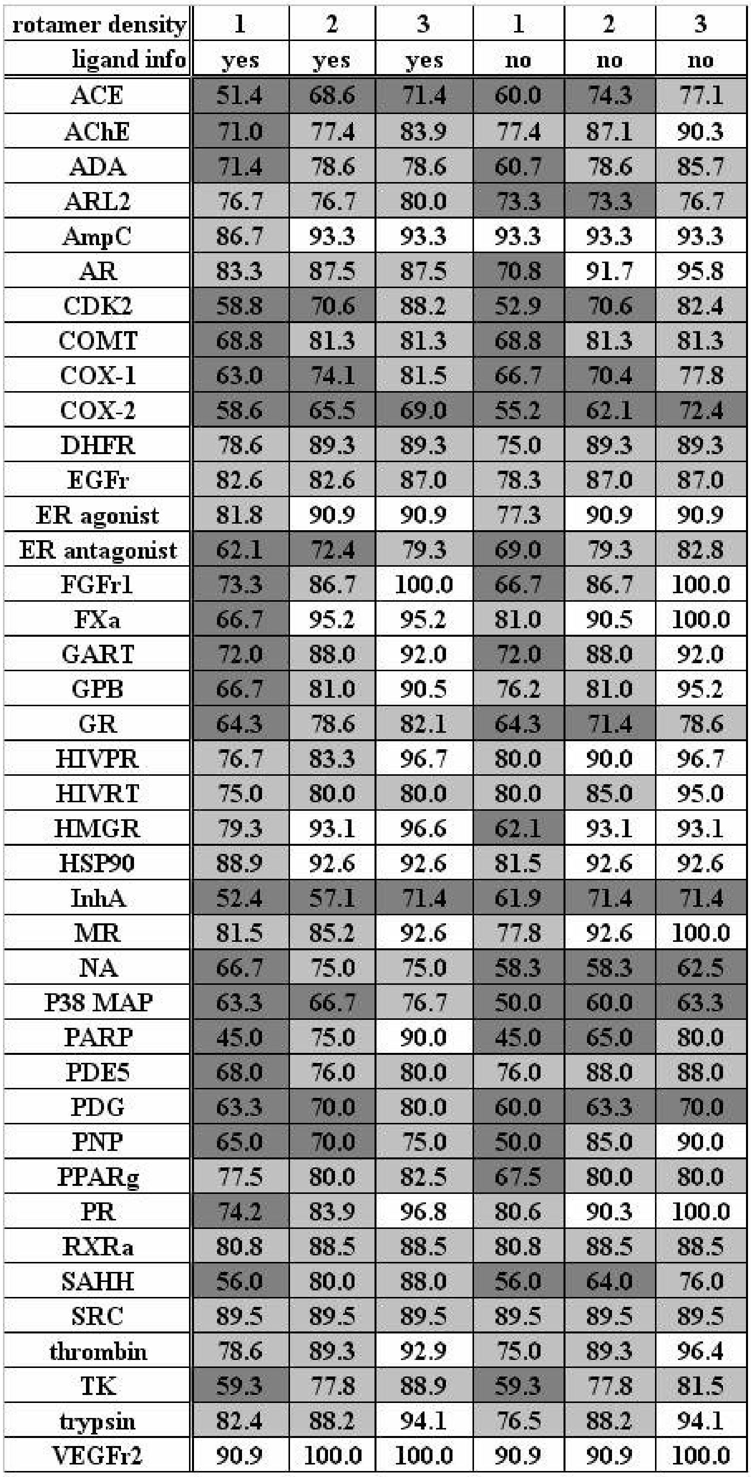
models still have an average atom displacement below the commonly accepted upper limit of 1.5 Å that a docking program can handle [7]. However, the small but inherent inaccuracy of the template X-ray structures must also be considered. A single side chain that points wrongly into the active site instead of pointing away can ruin a complete docking trial. Therefore it is often preferable to have multiple conformations for such side chains that allow for selecting a non-clashing variant. Table 6.3 lists the percentages of side chains for each individual IRECS model that have correct $\chi_{1,2}$ dihedral angles.

6.1.3 Step 3: Docking of Ligands with FlexE and FlexX

For all models the required receptor and ensemble definition files were build with default values (see FlexX manual [168]). In the case of the proteins ACE, ADA, COX-1, GR and HIVPR (see Table 6.1 for abbreviations) the surface computation required to manual define a single point that was located in the active site and outside the protein. Hydrogen positions were defined with the default FlexX torsion angles. The protonation of amino acids was also assigned with the default FlexX templates for the respective amino acids. Zink, magnesium, calcium and iron atoms are assigned preconfigured F-Score interaction templates, as well as the cofactors nicotinamide-adenine-dinucleotide phosphate (NAP) and dihydronicotinamide-adenine-dinucleotide phosphate (NDP). All other atoms of non-standard amino acids were removed from the active site, including the ligand and water molecules.

FlexX was used to dock into all rigid models, whereas FlexE was used to dock into models with rotamer density two and three. This section lists and describes all changes that were applied to the implementation and parameterization of FlexX and FlexE when docking ligands into proteins from the DUD.

1. According to the rotamer density (chosen as integer values), FlexE reads in that many multiple copies of the protein models, each time selecting a different set of rotamers for the side chains according to the alternative allocation identifiers in the respective PDB files. If for a side chain fewer rotamers are defined by IRECS than the total number



rotamer density	1	2	3	1	2	3
ligand info	yes	yes	yes	no	no	no
ACE	51.4	68.6	71.4	60.0	74.3	77.1
AChE	71.0	77.4	83.9	77.4	87.1	90.3
ADA	71.4	78.6	78.6	60.7	78.6	85.7
ARL2	76.7	76.7	80.0	73.3	73.3	76.7
AmpC	86.7	93.3	93.3	93.3	93.3	93.3
AR	83.3	87.5	87.5	70.8	91.7	95.8
CDK2	58.8	70.6	88.2	52.9	70.6	82.4
COMT	68.8	81.3	81.3	68.8	81.3	81.3
COX-1	63.0	74.1	81.5	66.7	70.4	77.8
COX-2	58.6	65.5	69.0	55.2	62.1	72.4
DHFR	78.6	89.3	89.3	75.0	89.3	89.3
EGFr	82.6	82.6	87.0	78.3	87.0	87.0
ER agonist	81.8	90.9	90.9	77.3	90.9	90.9
ER antagonist	62.1	72.4	79.3	69.0	79.3	82.8
FGFr1	73.3	86.7	100.0	66.7	86.7	100.0
FXa	66.7	95.2	95.2	81.0	90.5	100.0
GART	72.0	88.0	92.0	72.0	88.0	92.0
GPB	66.7	81.0	90.5	76.2	81.0	95.2
GR	64.3	78.6	82.1	64.3	71.4	78.6
HIVPR	76.7	83.3	96.7	80.0	90.0	96.7
HIVRT	75.0	80.0	80.0	80.0	85.0	95.0
HMGR	79.3	93.1	96.6	62.1	93.1	93.1
HSP90	88.9	92.6	92.6	81.5	92.6	92.6
InhA	52.4	57.1	71.4	61.9	71.4	71.4
MR	81.5	85.2	92.6	77.8	92.6	100.0
NA	66.7	75.0	75.0	58.3	58.3	62.5
P38 MAP	63.3	66.7	76.7	50.0	60.0	63.3
PARP	45.0	75.0	90.0	45.0	65.0	80.0
PDE5	68.0	76.0	80.0	76.0	88.0	88.0
PDG	63.3	70.0	80.0	60.0	63.3	70.0
PNP	65.0	70.0	75.0	50.0	85.0	90.0
PPARg	77.5	80.0	82.5	67.5	80.0	80.0
PR	74.2	83.9	96.8	80.6	90.3	100.0
RXRa	80.8	88.5	88.5	80.8	88.5	88.5
SAHH	56.0	80.0	88.0	56.0	64.0	76.0
SRC	89.5	89.5	89.5	89.5	89.5	89.5
thrombin	78.6	89.3	92.9	75.0	89.3	96.4
TK	59.3	77.8	88.9	59.3	77.8	81.5
trypsin	82.4	88.2	94.1	76.5	88.2	94.1
VEGFr2	90.9	100.0	100.0	90.9	90.9	100.0

Table 6.3: Percentage of side chains with correct $\chi_{1,2}$ dihedral angles in IRECS models. Field are shaded according to their value (dark - low values, light - high values).

of copies, an empty rotamer is assigned to that side chain. This allows for ignoring the interactions with this side chain (except C_β atoms) completely when docking with FlexE, which is a tribute to the low accuracy of side-chain prediction.

2. No superposition was required for building the united protein description of FlexE since all IRECS models of a protein use the backbone of this protein as template as defined in the respective X-ray structure.

3. Clustering was only performed for backbone and C_β atoms since the rotamers of the BBDep already represent clustered side-chain conformations. The united protein description of FlexE therefore contains only single conformations of the protein backbone. LUDI interaction points were also not clustered.

4. Logical and structural incompatibilities were computed, but no geometric incompatibility between rotamers or between rotamers and backbone were determined, since the rotamers can have slight clashes with the backbone or with each other. It is assumed here that such small clashes can be relaxed with an energy minimization procedure after complex construction if required and therefore usually do not lead to unrealistic protein conformations.

5. The number of solutions that are kept during each buildup step of the ligand was increased from 200 to 300. This is required since the conformational space resulting from the conformational degrees of freedom of protein *and* ligand is much larger for FlexE than for FlexX (only ligand flexibility) and therefore also requires more intermediate solutions to find a near-native complex conformation.

6. The internal factor for ligand atom clash checks 'CLASH_FACTOR' is reduced from 0.6 to 0.4. This change contributes to the fact that a ligand which tries to adapt to a disturbed protein conformation will sometimes also have to adopt a distorted conformation, which is facilitated by this more tolerant parameter setting.

7. FlexE uses the clique enumeration algorithm of Bron and Kerbosch [23] for instance selection. This algorithm is replaced by the Self-Consistent Mean Field (SCMF) algorithm [110] that was implemented into FlexE previously [71].

8. If ROTA is used for ligand scoring it completely replaces all F-Scores that usually guide the ligand build-up inside the active site and compute an estimate of the binding free energy. The ROTA version for docking is used here exclusively. Nevertheless, the LUDI interaction geometries still mainly determine the conformation sampling of the ligands, as the base fragments are still placed between surface points of these interaction geometries.

9. FlexE and FlexX score the placement solutions of partial ligands by (i) the currently achieved interaction score of the partial buildup ligand with the protein plus (ii) an estimate for the maximum score that the remaining fragments can achieve if they can establish optimal interactions with the protein [171]. This is required since FlexX uses different fragmentations and build-up orders simultaneously. If a certain partial

solution contains a high scoring fragment, it would easily supersede all other partial solutions that do not contain this fragment, yet. ROTA does not have this problem since there are no such score dominating features like hydrogen bonds. The score of a fragment is mainly dependent on the number of its atoms and thus the score of a partial ligand is divided by its atoms for estimating its maximum score.

6.1.4 Step 4: Evaluation of Docking Performance

The performance of FlexX and FlexE in redocking experiments is determined by comparing the generated conformations with the native ligand conformation found in the respective X-ray structure and calculating their pairwise RMSD. It is common to apply an RMSD cutoff of 2.0 Å to determine if a ligand conformation was predicted correctly or not. It is meaningful to increase this cutoff for docking into flexible proteins since the reference interaction geometry for the ligand also changes and thus a native-like binding mode can be assumed at larger RMSD cutoffs if both ligand and protein undertake medium scale (0.5 Å - 1.5 Å) conformational adaptations that result in a reconstruction of the experimental determined binding mode. Therefore, RMSD values were considered up to 3.0 Å. Often a docking program is able to generate near-native ligand conformations but fails to appropriate rank them . In these cases the application of another scoring function or a post-optimization of the complex conformation can improve the ranking and lift the near-native conformations to the top. Therefore it is also meaningful to benchmark a docking program with respect to its ability to generate a near-native ligand conformation at *any* rank.

The enrichment factor of a virtual screening experiment is computed as given in Huang et al. [85]: the enrichment factor EF_{subset} of a subset of the total ranked target-specific compound database is calculated by comparing the fraction of ligands among all compounds in this selected subset by the same fraction among all compounds of the whole target-specific database:

$$EF_{\text{subset}} = \frac{|ligands\ \cap\ subset|/|subset|}{|ligands|/|compounds|} \tag{6.1}$$

The ratio $|ligands|/|compounds|$ is always around 36, since the number of decoys is usually about the number of ligands times 35.

The enrichment factor is calculated for subsets comprising 1% (EF_1) and 20% (EF_{20}) of all compounds. These cutoff values are arbitrary selections but they are used commonly in realistic screening applications. The enrichment factor EF_{max} refers to the highest possible enrichment factor that can be achieved when selecting a subset of the database. Two aspects have to be considered when comparing enrichment factors of screening setups of different proteins. First, the compounds that are marked as decoys are putative non-binders and in general no tests were performed that actually showed that these compounds do not bind to the respective proteins [84]. Since the number of unidentified binders in the decoy sets is unknown and can vary between targets, the comparison of enrichment factors of screening setups for different targets may be biased towards proteins that have exceptional low numbers of unidentified ligands in their decoy sets. The second aspect to be considered is that the decoy compounds share different degrees of molecular similarity (topological, interaction profiles) with each other and with the ligands. This phenomenon can lead to different levels of difficulty in distinguishing ligands from decoys for the docking programs and must also be considered when enrichment factors of screening runs on different proteins are compared.

6.2 Evaluation: Docking into Protein Models Generated with IRECS

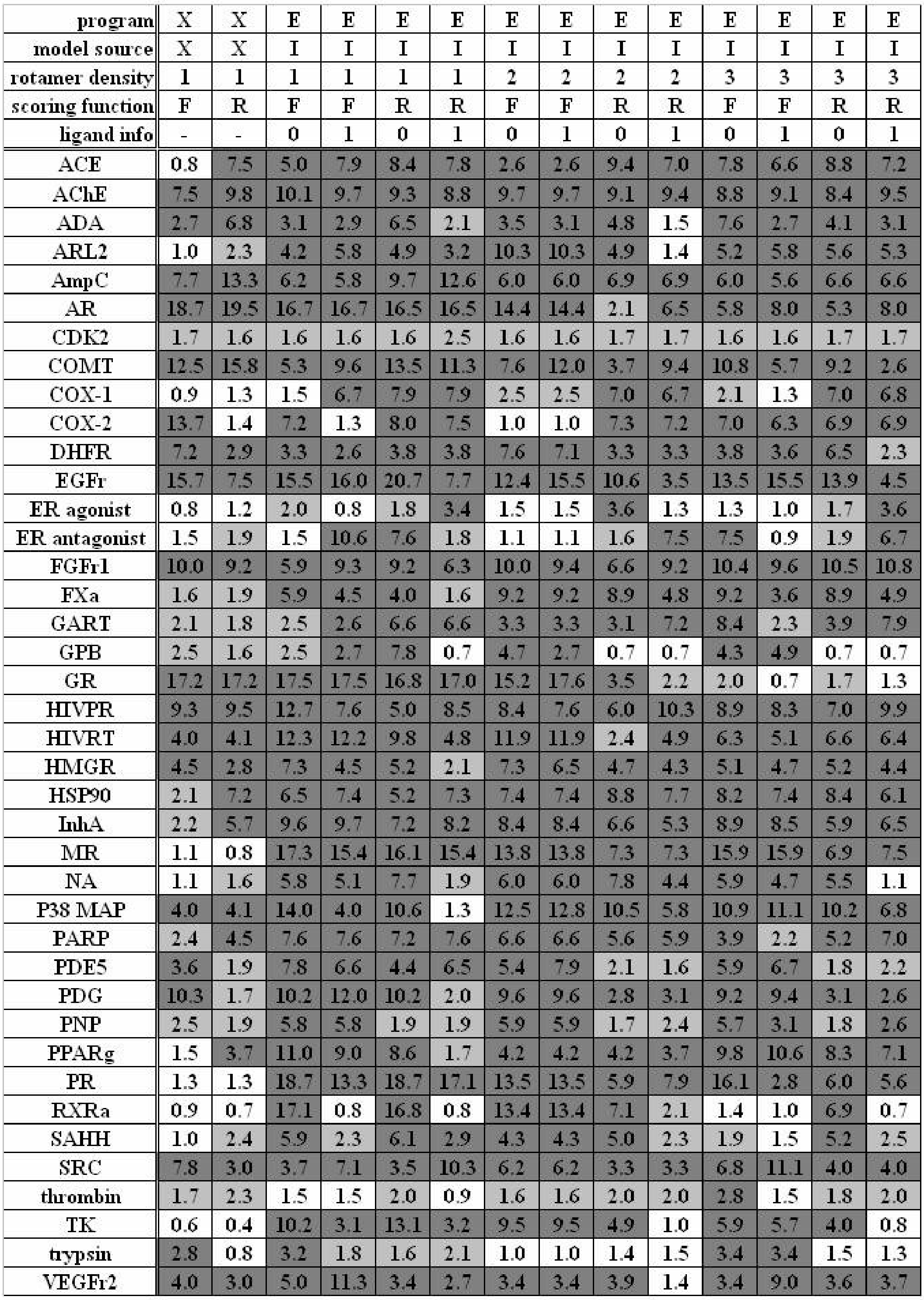
The evaluation of the proposed modeling and docking pipeline has the aim of showing (i) the advantages and disadvantages of using ROTA as a scoring function for generation of complex conformations and estimation of binding affinity, (ii) the influence of the inclusion of the native ligand conformation in the IRECS modeling procedure on the modeling and docking quality and (iii) the effects of handling the side-chain flexibility during docking. The evaluation consists of redocking and screening experiments on the forty protein targets of the DUD. These two kinds of experiments were chosen since they are standard comparison methods for docking programs and represent different challenges for the modeling and docking pipeline that are relevant for determining its usability in pharmaceutical research: the redocking experiments contain the task of comparing homogeneous complex conformations, where all ranked complexes consist of the same ligand and protein, whereas the screening experiments compare heterogeneous complex conformations, in which the protein remains the same but different ligands are bound.

6.2.1 Redocking using X-ray Structures and IRECS Models

Tables 6.4 and 6.5 show the RMSD values between generated complex conformations and the native conformations of the respective complexes. Table 6.4 shows the RMSD values of the top-ranked complex conformations and Table 6.5 shows the RMSD values of the generated complex conformations that are structurally closest to their respective experimental reference conformations among all generated conformations. Both tables show results for all reported setups for modeling (inclusion of ligand information, side-chain flexibility), scoring (ROTA or F-Score) and docking (FlexX or FlexE). Table 6.6 summarizes these results and allows for quickly comparing the performance of the different setups by counting for each experiment the amount of proteins for which a particular experimental setup was able to generate near-native complex conformations using multiple RMSD cutoffs.

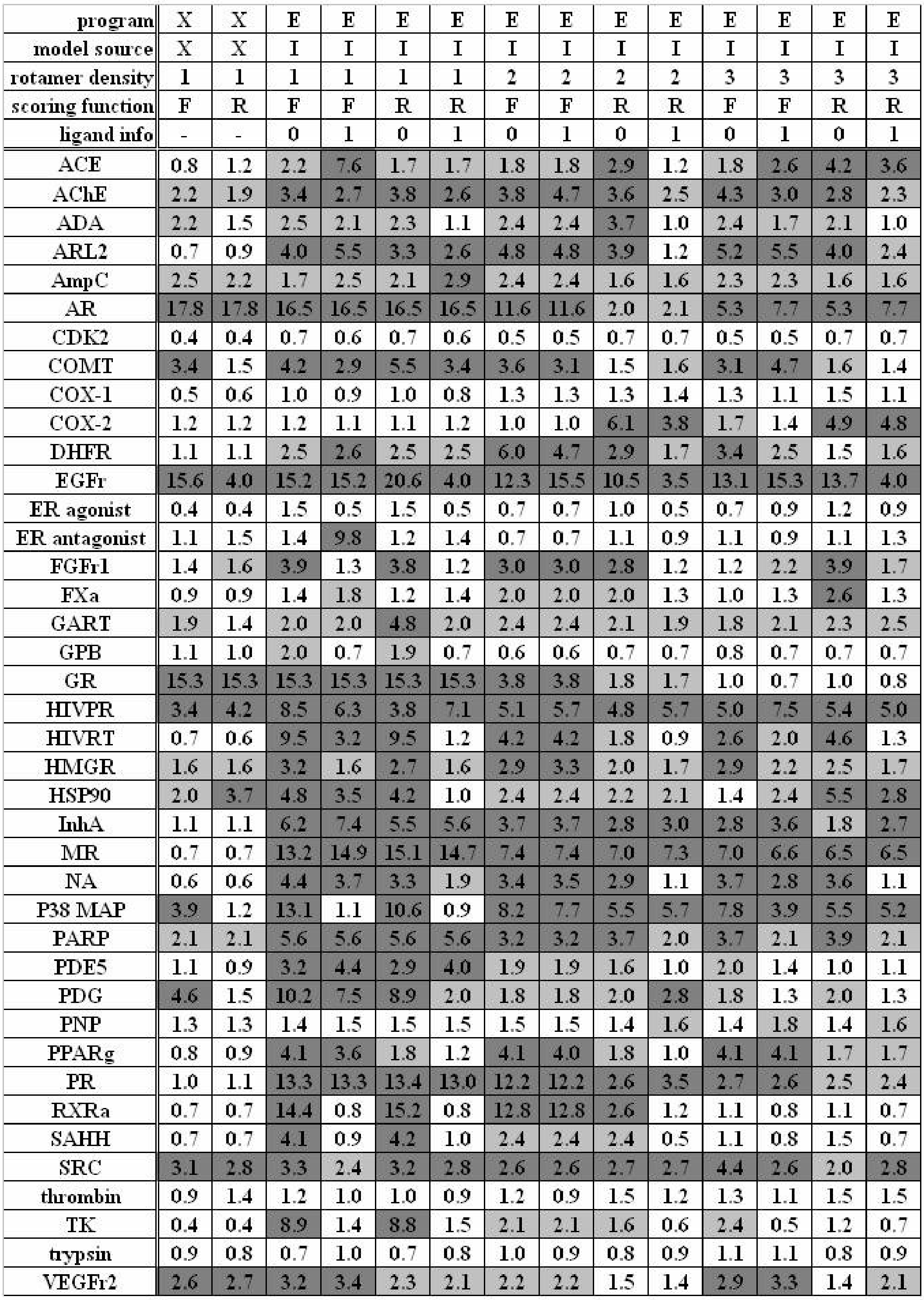
Effects of Ligand-Based Side-Chain Prediction

The inclusion of ligand information during IRECS optimization enables the ligand to imprint the conformational information of its bound conformation into the IRECS selection of rotamer ensembles, within the limitations of the BBDep rotamer library. This technique therefore first enables the side chains to adapt to the ligand bound mode and second to shape a binding pocket for bound compounds. Without this technique it is often observed that protein side chains modeled with IRECS fill empty cavities and contribute to intra-protein hydrogen bond networks. This modeling technique basically reproduces the apo-form of the protein structure and this form is rarely usable for accurately docking ligands. Table 6.2



program	X	X	E	E	E	E	E	E	E	E	E	E	E	E
model source	X	X	I	I	I	I	I	I	I	I	I	I	I	I
rotamer density	1	1	1	1	1	1	2	2	2	2	3	3	3	3
scoring function	F	R	F	F	R	R	F	F	R	R	F	F	R	R
ligand info	-	-	0	1	0	1	0	1	0	1	0	1	0	1
ACE	0.8	7.5	5.0	7.9	8.4	7.8	2.6	2.6	9.4	7.0	7.8	6.6	8.8	7.2
AChE	7.5	9.8	10.1	9.7	9.3	8.8	9.7	9.7	9.1	9.4	8.8	9.1	8.4	9.5
ADA	2.7	6.8	3.1	2.9	6.5	2.1	3.5	3.1	4.8	1.5	7.6	2.7	4.1	3.1
ARL2	1.0	2.3	4.2	5.8	4.9	3.2	10.3	10.3	4.9	1.4	5.2	5.8	5.6	5.3
AmpC	7.7	13.3	6.2	5.8	9.7	12.6	6.0	6.0	6.9	6.9	6.0	5.6	6.6	6.6
AR	18.7	19.5	16.7	16.7	16.5	16.5	14.4	14.4	2.1	6.5	5.8	8.0	5.3	8.0
CDK2	1.7	1.6	1.6	1.6	1.6	2.5	1.6	1.6	1.7	1.7	1.6	1.6	1.7	1.7
COMT	12.5	15.8	5.3	9.6	13.5	11.3	7.6	12.0	3.7	9.4	10.8	5.7	9.2	2.6
COX-1	0.9	1.3	1.5	6.7	7.9	7.9	2.5	2.5	7.0	6.7	2.1	1.3	7.0	6.8
COX-2	13.7	1.4	7.2	1.3	8.0	7.5	1.0	1.0	7.3	7.2	7.0	6.3	6.9	6.9
DHFR	7.2	2.9	3.3	2.6	3.8	3.8	7.6	7.1	3.3	3.3	3.8	3.6	6.5	2.3
EGFr	15.7	7.5	15.5	16.0	20.7	7.7	12.4	15.5	10.6	3.5	13.5	15.5	13.9	4.5
ER agonist	0.8	1.2	2.0	0.8	1.8	3.4	1.5	1.5	3.6	1.3	1.3	1.0	1.7	3.6
ER antagonist	1.5	1.9	1.5	10.6	7.6	1.8	1.1	1.1	1.6	7.5	7.5	0.9	1.9	6.7
FGFr1	10.0	9.2	5.9	9.3	9.2	6.3	10.0	9.4	6.6	9.2	10.4	9.6	10.5	10.8
FXa	1.6	1.9	5.9	4.5	4.0	1.6	9.2	9.2	8.9	4.8	9.2	3.6	8.9	4.9
GART	2.1	1.8	2.5	2.6	6.6	6.6	3.3	3.3	3.1	7.2	8.4	2.3	3.9	7.9
GPB	2.5	1.6	2.5	2.7	7.8	0.7	4.7	2.7	0.7	0.7	4.3	4.9	0.7	0.7
GR	17.2	17.2	17.5	17.5	16.8	17.0	15.2	17.6	3.5	2.2	2.0	0.7	1.7	1.3
HIVPR	9.3	9.5	12.7	7.6	5.0	8.5	8.4	7.6	6.0	10.3	8.9	8.3	7.0	9.9
HIVRT	4.0	4.1	12.3	12.2	9.8	4.8	11.9	11.9	2.4	4.9	6.3	5.1	6.6	6.4
HMGR	4.5	2.8	7.3	4.5	5.2	2.1	7.3	6.5	4.7	4.3	5.1	4.7	5.2	4.4
HSP90	2.1	7.2	6.5	7.4	5.2	7.3	7.4	7.4	8.8	7.7	8.2	7.4	8.4	6.1
InhA	2.2	5.7	9.6	9.7	7.2	8.2	8.4	8.4	6.6	5.3	8.9	8.5	5.9	6.5
MR	1.1	0.8	17.3	15.4	16.1	15.4	13.8	13.8	7.3	7.3	15.9	15.9	6.9	7.5
NA	1.1	1.6	5.8	5.1	7.7	1.9	6.0	6.0	7.8	4.4	5.9	4.7	5.5	1.1
P38 MAP	4.0	4.1	14.0	4.0	10.6	1.3	12.5	12.8	10.5	5.8	10.9	11.1	10.2	6.8
PARP	2.4	4.5	7.6	7.6	7.2	7.6	6.6	6.6	5.6	5.9	3.9	2.2	5.2	7.0
PDE5	3.6	1.9	7.8	6.6	4.4	6.5	5.4	7.9	2.1	1.6	5.9	6.7	1.8	2.2
PDG	10.3	1.7	10.2	12.0	10.2	2.0	9.6	9.6	2.8	3.1	9.2	9.4	3.1	2.6
PNP	2.5	1.9	5.8	5.8	1.9	1.9	5.9	5.9	1.7	2.4	5.7	3.1	1.8	2.6
PPARg	1.5	3.7	11.0	9.0	8.6	1.7	4.2	4.2	4.2	3.7	9.8	10.6	8.3	7.1
PR	1.3	1.3	18.7	13.3	18.7	17.1	13.5	13.5	5.9	7.9	16.1	2.8	6.0	5.6
RXRa	0.9	0.7	17.1	0.8	16.8	0.8	13.4	13.4	7.1	2.1	1.4	1.0	6.9	0.7
SAHH	1.0	2.4	5.9	2.3	6.1	2.9	4.3	4.3	5.0	2.3	1.9	1.5	5.2	2.5
SRC	7.8	3.0	3.7	7.1	3.5	10.3	6.2	6.2	3.3	3.3	6.8	11.1	4.0	4.0
thrombin	1.7	2.3	1.5	1.5	2.0	0.9	1.6	1.6	2.0	2.0	2.8	1.5	1.8	2.0
TK	0.6	0.4	10.2	3.1	13.1	3.2	9.5	9.5	4.9	1.0	5.9	5.7	4.0	0.8
trypsin	2.8	0.8	3.2	1.8	1.6	2.1	1.0	1.0	1.4	1.5	3.4	3.4	1.5	1.3
VEGFr2	4.0	3.0	5.0	11.3	3.4	2.7	3.4	3.4	3.9	1.4	3.4	9.0	3.6	3.7

Table 6.4: RMSD values of the top-ranked complex conformations as compared to the native conformations. Program: X = FlexX, E = FlexE; model source: X = X-ray, I = IRECS; scoring function: F = F-Score, R = ROTA. Field are shaded according to their value (dark - high values, light - low values).



program	X	X	E	E	E	E	E	E	E	E	E	E	E	E
model source	X	X	I	I	I	I	I	I	I	I	I	I	I	I
rotamer density	1	1	1	1	1	1	2	2	2	2	3	3	3	3
scoring function	F	R	F	F	R	R	F	F	R	R	F	F	R	R
ligand info	-	-	0	1	0	1	0	1	0	1	0	1	0	1
ACE	0.8	1.2	2.2	7.6	1.7	1.7	1.8	1.8	2.9	1.2	1.8	2.6	4.2	3.6
AChE	2.2	1.9	3.4	2.7	3.8	2.6	3.8	4.7	3.6	2.5	4.3	3.0	2.8	2.3
ADA	2.2	1.5	2.5	2.1	2.3	1.1	2.4	2.4	3.7	1.0	2.4	1.7	2.1	1.0
ARL2	0.7	0.9	4.0	5.5	3.3	2.6	4.8	4.8	3.9	1.2	5.2	5.5	4.0	2.4
AmpC	2.5	2.2	1.7	2.5	2.1	2.9	2.4	2.4	1.6	1.6	2.3	2.3	1.6	1.6
AR	17.8	17.8	16.5	16.5	16.5	16.5	11.6	11.6	2.0	2.1	5.3	7.7	5.3	7.7
CDK2	0.4	0.4	0.7	0.6	0.7	0.6	0.5	0.5	0.7	0.7	0.5	0.5	0.7	0.7
COMT	3.4	1.5	4.2	2.9	5.5	3.4	3.6	3.1	1.5	1.6	3.1	4.7	1.6	1.4
COX-1	0.5	0.6	1.0	0.9	1.0	0.8	1.3	1.3	1.3	1.4	1.3	1.1	1.5	1.1
COX-2	1.2	1.2	1.2	1.1	1.1	1.2	1.0	1.0	6.1	3.8	1.7	1.4	4.9	4.8
DHFR	1.1	1.1	2.5	2.6	2.5	2.5	6.0	4.7	2.9	1.7	3.4	2.5	1.5	1.6
EGFr	15.6	4.0	15.2	15.2	20.6	4.0	12.3	15.5	10.5	3.5	13.1	15.3	13.7	4.0
ER agonist	0.4	0.4	1.5	0.5	1.5	0.5	0.7	0.7	1.0	0.5	0.7	0.9	1.2	0.9
ER antagonist	1.1	1.5	1.4	9.8	1.2	1.4	0.7	0.7	1.1	0.9	1.1	0.9	1.1	1.3
FGFr1	1.4	1.6	3.9	1.3	3.8	1.2	3.0	3.0	2.8	1.2	1.2	2.2	3.9	1.7
FXa	0.9	0.9	1.4	1.8	1.2	1.4	2.0	2.0	2.0	1.3	1.0	1.3	2.6	1.3
GART	1.9	1.4	2.0	2.0	4.8	2.0	2.4	2.4	2.1	1.9	1.8	2.1	2.3	2.5
GPB	1.1	1.0	2.0	0.7	1.9	0.7	0.6	0.6	0.7	0.7	0.8	0.7	0.7	0.7
GR	15.3	15.3	15.3	15.3	15.3	15.3	3.8	3.8	1.8	1.7	1.0	0.7	1.0	0.8
HIVPR	3.4	4.2	8.5	6.3	3.8	7.1	5.1	5.7	4.8	5.7	5.0	7.5	5.4	5.0
HIVRT	0.7	0.6	9.5	3.2	9.5	1.2	4.2	4.2	1.8	0.9	2.6	2.0	4.6	1.3
HMGR	1.6	1.6	3.2	1.6	2.7	1.6	2.9	3.3	2.0	1.7	2.9	2.2	2.5	1.7
HSP90	2.0	3.7	4.8	3.5	4.2	1.0	2.4	2.4	2.2	2.1	1.4	2.4	5.5	2.8
InhA	1.1	1.1	6.2	7.4	5.5	5.6	3.7	3.7	2.8	3.0	2.8	3.6	1.8	2.7
MR	0.7	0.7	13.2	14.9	15.1	14.7	7.4	7.4	7.0	7.3	7.0	6.6	6.5	6.5
NA	0.6	0.6	4.4	3.7	3.3	1.9	3.4	3.5	2.9	1.1	3.7	2.8	3.6	1.1
P38 MAP	3.9	1.2	13.1	1.1	10.6	0.9	8.2	7.7	5.5	5.7	7.8	3.9	5.5	5.2
PARP	2.1	2.1	5.6	5.6	5.6	5.6	3.2	3.2	3.7	2.0	3.7	2.1	3.9	2.1
PDE5	1.1	0.9	3.2	4.4	2.9	4.0	1.9	1.9	1.6	1.0	2.0	1.4	1.0	1.1
PDG	4.6	1.5	10.2	7.5	8.9	2.0	1.8	1.8	2.0	2.8	1.8	1.3	2.0	1.3
PNP	1.3	1.3	1.4	1.5	1.5	1.5	1.5	1.5	1.4	1.6	1.4	1.8	1.4	1.6
PPARg	0.8	0.9	4.1	3.6	1.8	1.2	4.1	4.0	1.8	1.0	4.1	4.1	1.7	1.7
PR	1.0	1.1	13.3	13.3	13.4	13.0	12.2	12.2	2.6	3.5	2.7	2.6	2.5	2.4
RXRa	0.7	0.7	14.4	0.8	15.2	0.8	12.8	12.8	2.6	1.2	1.1	0.8	1.1	0.7
SAHH	0.7	0.7	4.1	0.9	4.2	1.0	2.4	2.4	2.4	0.5	1.1	0.8	1.5	0.7
SRC	3.1	2.8	3.3	2.4	3.2	2.8	2.6	2.6	2.7	2.7	4.4	2.6	2.0	2.8
thrombin	0.9	1.4	1.2	1.0	1.0	0.9	1.2	0.9	1.5	1.2	1.3	1.1	1.5	1.5
TK	0.4	0.4	8.9	1.4	8.8	1.5	2.1	2.1	1.6	0.6	2.4	0.5	1.2	0.7
trypsin	0.9	0.8	0.7	1.0	0.7	0.8	1.0	0.9	0.8	0.9	1.1	1.1	0.8	0.9
VEGFr2	2.6	2.7	3.2	3.4	2.3	2.1	2.2	2.2	1.5	1.4	2.9	3.3	1.4	2.1

Table 6.5: Minimum RMSD values among all generated complex conformations as compared to the native conformations. Program: X = FlexX, E = FlexE; model source: X = X-ray, I = IRECS; scoring function: F = F-Score, R = ROTA. Field are shaded according to their value (dark - high values, light - low values).

Table 6.6: Redocking results of FlexX and FlexE on IRECS models considering rank one

(a) Considering the first rank

	Model	Rotamer	Scoring	Ligand	RMSD ≤					
Tool	source	density	function	info	1.0	1.5	2.0	2.5	3.0	inf
FlexX	X-ray	1	F-Score	yes	7	11	15	20	23	40
FlexX	X-ray	1	ROTA	yes	4	8	17	20	23	40
FlexX	IRECS	1	F-Score	no	0	3	5	6	7	40
FlexX	IRECS	1	F-Score	yes	2	4	6	7	11	40
FlexX	IRECS	1	ROTA	no	0	0	5	5	5	40
FlexX	IRECS	1	ROTA	yes	3	4	10	13	16	40
FlexE	IRECS	2	F-Score	no	0	4	6	6	8	40
FlexE	IRECS	2	F-Score	yes	0	4	6	6	9	40
FlexE	IRECS	2	ROTA	no	1	2	6	9	10	40
FlexE	IRECS	2	ROTA	yes	2	6	9	14	14	40
FlexE	IRECS	3	F-Score	no	0	2	4	6	7	40
FlexE	IRECS	3	F-Score	yes	3	7	8	10	12	40
FlexE	IRECS	3	ROTA	no	1	1	9	9	9	40
FlexE	IRECS	3	ROTA	yes	3	6	7	11	14	40

(b) Considering all ranks

	Model	Rotamer	Scoring	Ligand	RMSD ≤					
Tool	source	density	function	info	1.0	1.5	2.0	2.5	3.0	inf
FlexX	X-ray	1	F-Score	yes	15	24	27	31	32	40
FlexX	X-ray	1	ROTA	yes	15	27	31	33	35	40
FlexX	IRECS	1	F-Score	no	2	9	11	15	15	40
FlexX	IRECS	1	F-Score	yes	7	13	15	18	22	40
FlexX	IRECS	1	ROTA	no	2	9	12	15	18	40
FlexX	IRECS	1	ROTA	yes	10	17	24	26	30	40
FlexE	IRECS	2	F-Score	no	4	8	13	20	22	40
FlexE	IRECS	2	F-Score	yes	6	8	13	20	21	40
FlexE	IRECS	2	ROTA	no	4	7	20	23	31	40
FlexE	IRECS	2	ROTA	yes	10	20	27	30	34	40
FlexE	IRECS	3	F-Score	no	3	14	19	22	27	40
FlexE	IRECS	3	F-Score	yes	8	15	18	25	29	40
FlexE	IRECS	3	ROTA	no	3	14	21	24	27	40
FlexE	IRECS	3	ROTA	yes	8	17	24	30	33	40

shows the RMSDs of predicted side chains inside the active sites, which generally decrease if ligand information is used during the IRECS optimization. The same accuracy gain can be observed when comparing the χ_1 accuracy of modeled side chains, which increases when a ligand is used during optimization. However, the $\chi_{1,2}$ accuracies slightly decrease. This seemingly contradictory result can be explained with the same arguments that were used to explain similar observations in Section 4.5.2. The inclusion of a ligand conformation during side-chain optimization leads to a much denser packing of atoms inside the active site, which makes it much harder to find the correct packing of side chains which are large but also have only few rotamers in their ensembles, like tryptophane and phenylalanine. The different accuracies for modeling side chains of different amino acids (see Figure 4.4) lead to the observed inconsistency.

The comparison of the results for the different experimental setups in Tables 6.4 and 6.5 and in Table 6.6 show that docking into protein models that were built using ligand information is more successful than docking into those models that were built without using ligand information. One can also see that this effect becomes weaker as the rotamer density increases. The reason for the latter observation is that the multiple rotamers of these IRECS models compensate for the above mentioned modeling problems. The advantage of including the ligand information in the modeling procedure is stronger than when considering all generated conformations when only the top-ranked conformation is considered.

Performance of F-Score and ROTA in Ranking Homogeneous Complex Conformations

All redocking experiments were performed both with ROTA and F-Score to respectively score ligand poses during the FlexX complex construction and final scoring. As one can see in Table 6.6 (a) F-Score is able to rank more native complex conformations at the top rank than ROTA when using X-ray structures. ROTA has a somewhat higher performance than F-Score if IRECS models are used. Both results were expected as ROTA is specially designed for scoring IRECS models and F-Score was designed to score X-ray structures. Table 6.6 (b) depicts the general ability of both scoring functions in guiding FlexX and FlexE towards near-native complex conformations. Here ROTA and F-Score perform similarly on the X-ray structures, but ROTA outperforms F-Score by a large margin if IRECS models are used for docking. The advantage of ROTA over F-Score more than doubles when docking into IRECS models with a rotamer density of two: FlexE can generate near-native complex conformations with a 2.0 Å cutoff for 27 proteins when ROTA is used for scoring, whereas FlexE can only generate such conformations for 13 proteins when using F-Score.

Influence of Side-Chain Flexibility on Redocking Accuracy

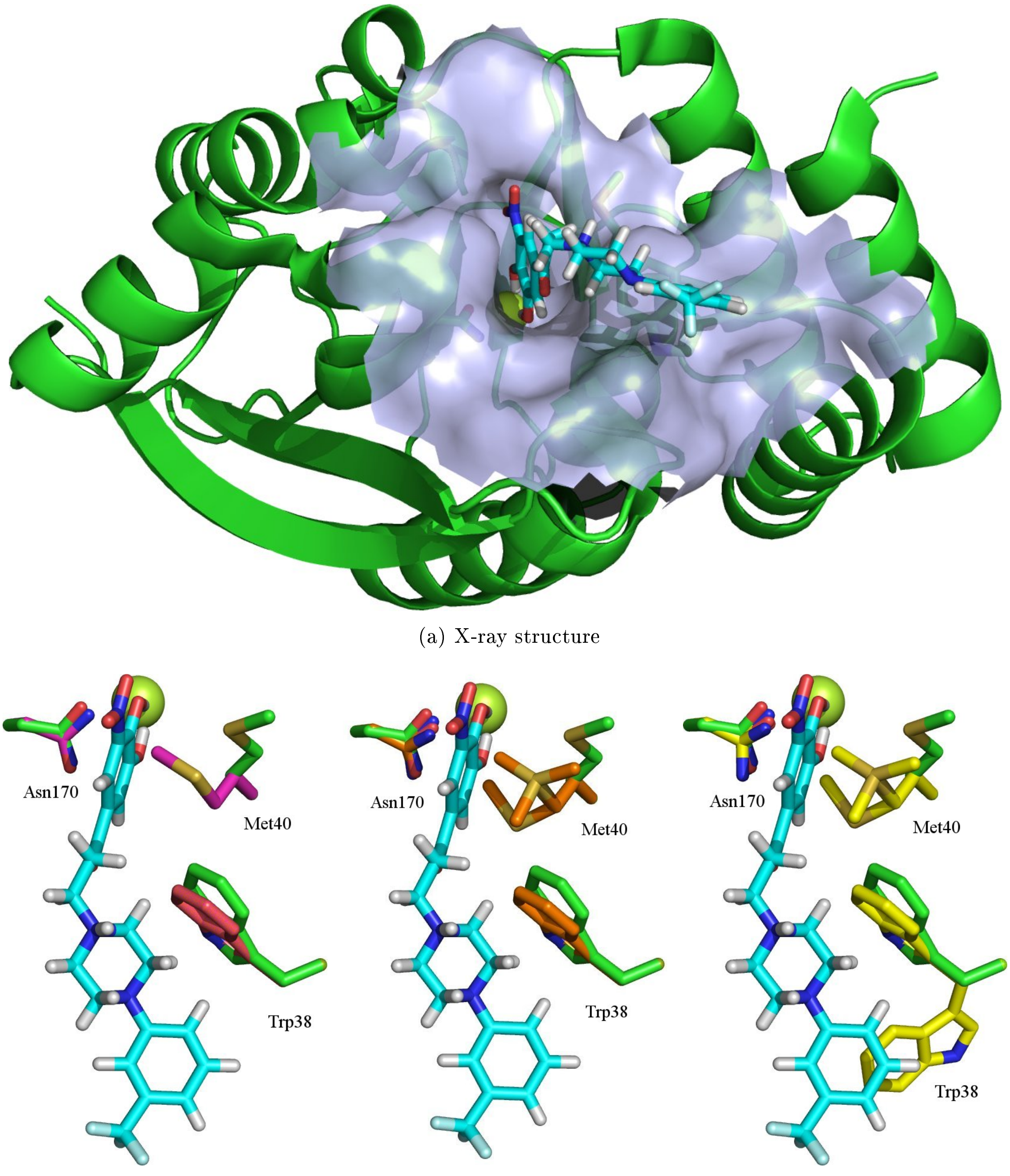
As stated in Section 6.2.1 X-ray structures of protein-ligand conformations are the best available protein models for redocking experiments. The usage of flexible protein models for redocking generally make docking harder since now the docking programs must both generate a near-native ligand conformation *and* a near-native protein conformation, and both depend on each other. The redocking accuracy is highest if FlexX is applied to X-ray structures. The accuracy of FlexE comes closest to this accuracy when scoring with ROTA and docking into IRECS models that were build with ligand information and two rotamers per side chain on average.

Among those redocking experiments that were run with IRECS models, the flexible models clearly outperform the rigid models. Here, the flexible models enable the docking program to select from a larger set of rotamers and allow for a much more accurate modeling of the protein during docking than it is the case for those models created with single rotamers per side chain. A drastic example that justifies the usage of protein flexibility is the protein GR, for which FlexX cannot create a near-native conformation but FlexE can do so by using flexible IRECS models. ALR2, COMT, DHFR, PARP and PDE5 are examples for proteins where the docking is only successful when using X-ray structures or flexible IRECS models, but using rigid IRECS models let to insufficient redocking accuracy. In contrast, for the protein MR IRECS is not able to generate a useful flexible model. FGFr1 is the only protein for which IRECS created a rigid model that is significantly better suited for docking than any flexible IRECS model (see Figure 6.5 and Section 6.2.2 for a structural explanation).

Redocking into Protein Models of the Catechol O-Methyltransferase

COMT methylates and inactivates L-DOPA which is used for treatment of Parkinson disease, and therefore COMT itself is a target for inhibition by pharmaceutical compounds [19]. This target is of special interest since the redocking experiments performed with ROTA have a much better performance than F-Score. Also, the rigid IRECS model has insufficience accuracy, wheras FlexE can find a ligand conformation with an RMSD below 2.0 Å. Figure 6.2(a) shows the catechol O-methyltransferase (COMT) from rat in complex with the inhibitor BIA 3-335 (PDB ID: 1h1d). The head group of BIA 3-335 binds tightly to the buried magnesium ion of COMT and to the polar side chains of Glu199, Asp141 and Asn170 via hydrogen bonds, the hydrophobic tail of the ligand just orients itself towards a hydrophobic patch of the protein surface. One general reason for the better performance of ROTA on this target is that F-Score is more dependent on hydrogen bonds than ROTA, which are only established between the protein and this head group.

Although FlexE is able to generate near-native complex conformations when using flexible IRECS models ($\rho_{rot} = 2$ and $\rho_{rot} = 3$) and ROTA for scoring (see Figure 6.6), the first ligand conformations with an RMSD below 2.0 Å to the conformation of the crystallized ligand are found on rank 59 ($\rho_{rot} = 2$) and on rank 39 ($\rho_{rot} = 3$). The higher ranked solutions have either the ligand tail shifted across the protein surface with the head bound to the Magnesium ion or the hydrophobic tail is docked inside the cavity, with the polar head pointing outward in the solvent. Figures 6.2(b), 6.2(c) and 6.2(d) show the three most important side chains of the IRECS models of COMT with one, two and three rotamers on average per side chain, respectively. The X-ray structure and its side chains are colored green and all models were generated without native ligand information. It can be seen that in the IRECS model with $\rho_{rot} = 1$ the tip of Met40 (upper right) occupies the same space as the ligand head. Therefore FlexX is unable to dock the ligand back to the binding pocket in a native-like conformation with the IRECS model. IRECS also wrongly predicts the χ_2 dihedral angle of Asn170 (upper left) which swaps hydrogen acceptor and donor function



(a) X-ray structure

(b) IRECS model, $\rho_{rot} = 1$

(c) IRECS model, $\rho_{rot} = 2$

(d) IRECS model, $\rho_{rot} = 3$

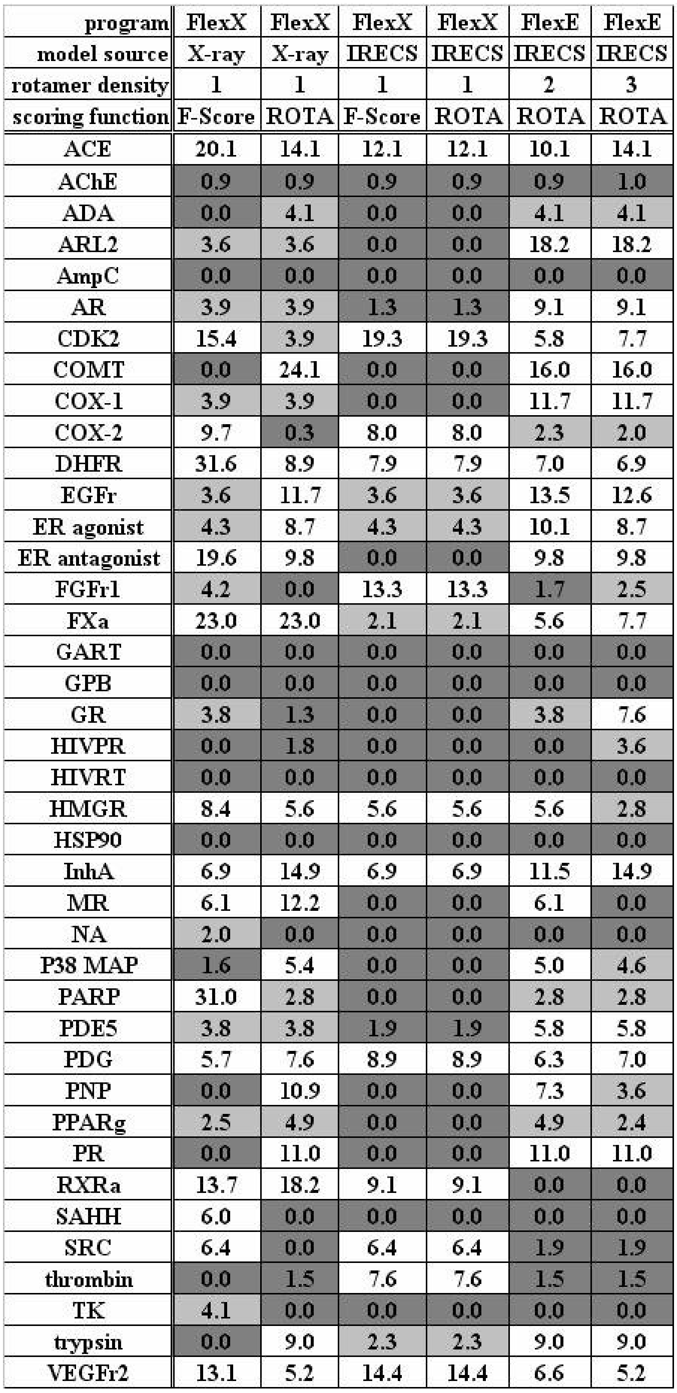
Figure 6.2: Side chains in the active site of COMT

and prevents the formation of a hydrogen bond between this side chain and the ligand head. Both flexible IRECS models have alternative conformations for these two side chains that enable the accurate docking of the ligand. Met40 is assigned three alternative conformations that are different to the conformations found in the X-ray structure, but they all do not clash with the native ligand conformation. Asn170 is assigned many rotamers with alternative χ_2 dihedral angles from which FlexE can select the appropriate one for hydrogen bonding. The IRECS model with three rotamers per side chain on average has an additional rotamer at Trp38 (lower right) that does not have influence on the docking run (as other side chains that were not drawn).

The results show that for this special target the combination of ROTA and flexible IRECS models was successful. This target makes a good example for the observation that for docking with FlexE it is more important that the rotamers complementary to the bound ligand are available than to exclude unadapted rotamers. This will become even clearer in the virtual screening experiments, where for each ligand there can be different adapted rotamers.

6.2.2 Screening of the Target-Specific Databases for Active Compounds

This section describes the results of the screening experiments that were performed using FlexX and FlexE, X-ray structures and IRECS models, different degrees of side-chain flexibility and the scoring functions F-Score and ROTA. Since in this test scenario the ligands binding to the proteins are known beforehand it is possible to draw an enrichment curve for each screening experiment. In an enrichment curve the percentage of retrieved ligands among any subset of a ranked database of compounds is measured. In a real screening scenario a certain percentage of the ranked database would be selected for further testing. The size of this subset usually depends on the size of the database, the expected number of ligands in the database, the expected quality of the ranking and the available resources for performing further tests. The curves that were derived with the experiments performed here can be found in appendix (see figures A.1 to A.14). Table 6.7 shows the achieved enrichment factors (see Equation 6.1) of the different screening experiments for all DUD targets that were achieved for the top-ranked 1% of the target-specific compound databases. From the table it can be seen that there is no single docking strategy that consistently displays a high screening performance: there are proteins for which all screening setups perform equally well (ACE, PDG) or badly (AChE, GPB, HIVRT, HSP90), whereas docking with side-chain flexibility is advantageous (ALR2, AR, COX-1, GR) or disadvantageous (GART, RXRa) for the screening of other proteins. There are also proteins for which the use of ROTA turned out to be advantageous (COMT, EGFr, PR) or disadvantageous (Thrombin). Furthermore, for some proteins (NA, SAHH, TK) FlexX cannot tolerate errors of the atom coordinates at all and therefore ligands for such proteins can only be retrieved successfully by using the respective X-ray structure. For a subset of the proteins similar screening results were previously published. For example, the observed low capacity of FlexX and FlexE using ROTA in achieving an enrichment of ligands in the compound database used for screening the neuraminidase seems to be a general flaw of knowledge-based scoring functions, since Stahl et al. [200] also reported that a screening with FlexX using the scoring function PLP



program	FlexX	FlexX	FlexX	FlexX	FlexE	FlexE
model source	X-ray	X-ray	IRECS	IRECS	IRECS	IRECS
rotamer density	1	1	1	1	2	3
scoring function	F-Score	ROTA	F-Score	ROTA	ROTA	ROTA
ACE	20.1	14.1	12.1	12.1	10.1	14.1
AChE	0.9	0.9	0.9	0.9	0.9	1.0
ADA	0.0	4.1	0.0	0.0	4.1	4.1
ARL2	3.6	3.6	0.0	0.0	18.2	18.2
AmpC	0.0	0.0	0.0	0.0	0.0	0.0
AR	3.9	3.9	1.3	1.3	9.1	9.1
CDK2	15.4	3.9	19.3	19.3	5.8	7.7
COMT	0.0	24.1	0.0	0.0	16.0	16.0
COX-1	3.9	3.9	0.0	0.0	11.7	11.7
COX-2	9.7	0.3	8.0	8.0	2.3	2.0
DHFR	31.6	8.9	7.9	7.9	7.0	6.9
EGFr	3.6	11.7	3.6	3.6	13.5	12.6
ER agonist	4.3	8.7	4.3	4.3	10.1	8.7
ER antagonist	19.6	9.8	0.0	0.0	9.8	9.8
FGFr1	4.2	0.0	13.3	13.3	1.7	2.5
FXa	23.0	23.0	2.1	2.1	5.6	7.7
GART	0.0	0.0	0.0	0.0	0.0	0.0
GPB	0.0	0.0	0.0	0.0	0.0	0.0
GR	3.8	1.3	0.0	0.0	3.8	7.6
HIVPR	0.0	1.8	0.0	0.0	0.0	3.6
HIVRT	0.0	0.0	0.0	0.0	0.0	0.0
HMGR	8.4	5.6	5.6	5.6	5.6	2.8
HSP90	0.0	0.0	0.0	0.0	0.0	0.0
InhA	6.9	14.9	6.9	6.9	11.5	14.9
MR	6.1	12.2	0.0	0.0	6.1	0.0
NA	2.0	0.0	0.0	0.0	0.0	0.0
P38 MAP	1.6	5.4	0.0	0.0	5.0	4.6
PARP	31.0	2.8	0.0	0.0	2.8	2.8
PDE5	3.8	3.8	1.9	1.9	5.8	5.8
PDG	5.7	7.6	8.9	8.9	6.3	7.0
PNP	0.0	10.9	0.0	0.0	7.3	3.6
PPARg	2.5	4.9	0.0	0.0	4.9	2.4
PR	0.0	11.0	0.0	0.0	11.0	11.0
RXRa	13.7	18.2	9.1	9.1	0.0	0.0
SAHH	6.0	0.0	0.0	0.0	0.0	0.0
SRC	6.4	0.0	6.4	6.4	1.9	1.9
thrombin	0.0	1.5	7.6	7.6	1.5	1.5
TK	4.1	0.0	0.0	0.0	0.0	0.0
trypsin	0.0	9.0	2.3	2.3	9.0	9.0
VEGFr2	13.1	5.2	14.4	14.4	6.6	5.2

Table 6.7: Enrichment factors for the DUD targets of the top 1% ranked databases. Program: X = FlexX, E = FlexE; model source: X = X-ray, I = IRECS; scoring function: F = F-Score, R = ROTA. Only the top 1% of the ranked databases were considered. Field are shaded according to their value (dark - low values, light - high values).

[60] yielded a much smaller enrichment factor compared to the factor that was achieved when using F-Score. The enrichment factors achieved here are generally somewhat lower than those of other screening studies with FlexX [200], FlexE [162] and other docking programs [101, 69, 145, 224]. The primary reason for this is the special compilation of the DUD target-specific databases, which intentionally poses much harder challenges for scoring than common compilations of screening databases [85].

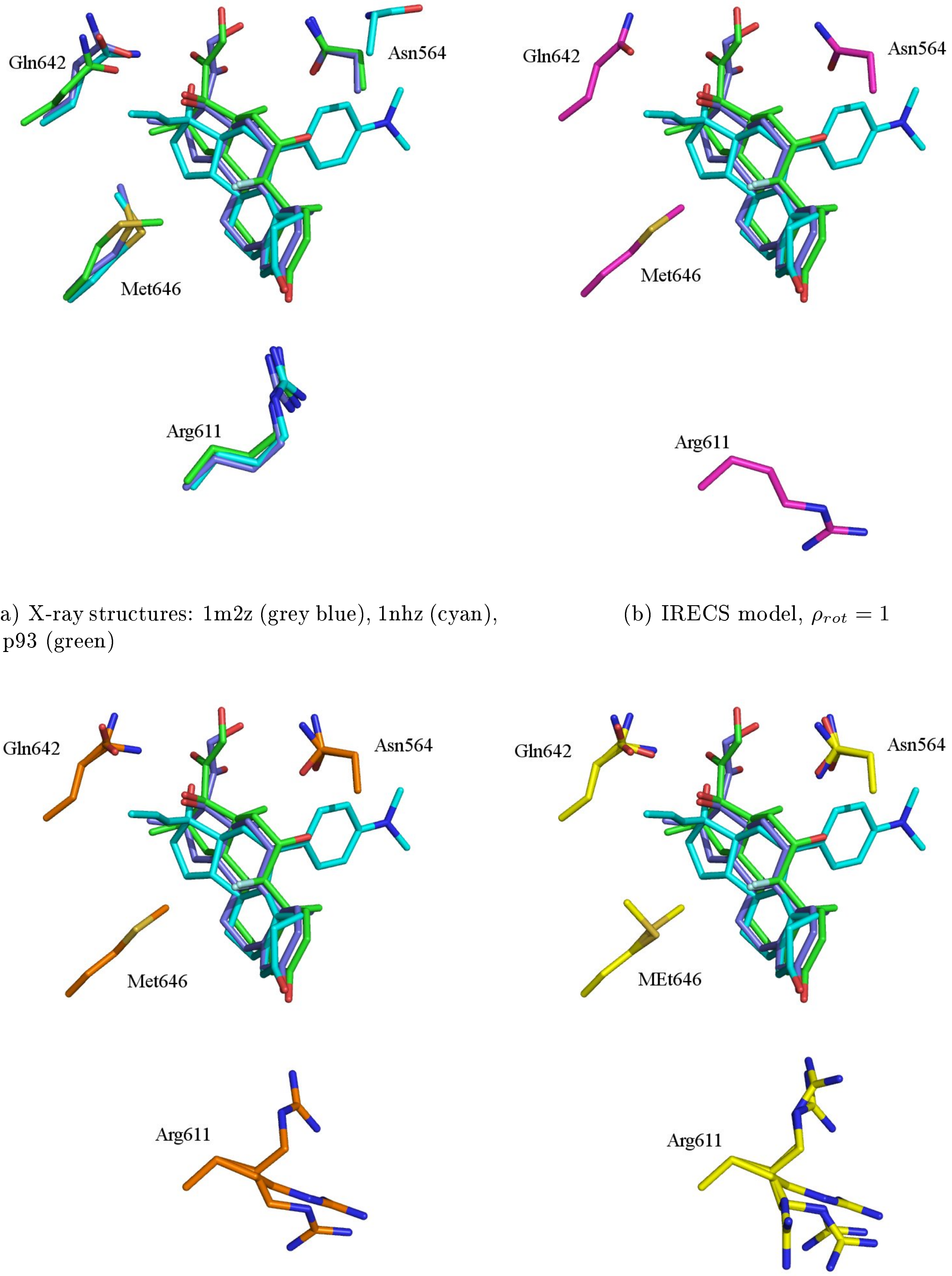
Performance of F-Score and ROTA in Ranking Heterogeneous Complexes

The enrichment factors shown in Figure 6.7 were derived from two screening setups that used F-Score and four screening setups that used ROTA for binding affinity prediction. All rigid protein models (X-ray structures and rigid IRECS models) were used for screening using F-Score and ROTA, which renders these setups suitable for comparing the performance of these two scoring functions. The setups performed with flexible models were performed to estimate the influence of side-chain flexibility on the overall screening performance, and thus no different scoring functions were tried out on these models. Additionally, from the redocking experiments it was known beforehand that all IRECS models independent from their flexibility perform best with the ROTA scoring function. Table 6.8 summarizes the enrichment factors for different subsets of the ranked compound databases (1%, 20%, best top selection) and the different screening setups. It can be seen that F-Score performs better than ROTA when an X-ray structure is used for docking and ROTA performs better than F-Score when an IRECS model is used for docking. This was expected as these functions were trained for their specific application purpose.

Influence of Side-Chain Flexibility on the Performance of Virtual Screening

The flexibility of the IRECS models enables FlexE to adapt the side-chain conformations to any partial or complete ligand conformation within narrow conformational limits. Compared to the traditional sampling and scoring scheme of FlexX (X-ray structures and F-Score) this strategy generally gives all compounds a higher chance of being docked by FlexE in an energetically favorable conformation. Since this is true for all docked compounds, active and inactive, this strategy does not automatically increase the performance of screening experiments.

When comparing the enrichment factors from Table 6.8, one can see that the performance of screening is nearly equal for those screening setups using IRECS models with $\rho_{rot} = 2$ and $\rho_{rot} = 3$. The only enrichment plots which exhibit different results for these setups are the enrichment plot for the $\text{ER}_{agonist}$ in Figure A.5 (better performance for $\rho_{rot} = 2$) and the enrichment plot for GR in Figure A.7 (better performance for $\rho_{rot} = 3$). This effect can be interpreted with the help of Figure 6.3, which depicts different structural models of the active site of GR: figure 6.3(a) shows a superposition of ligands and side chains of residues that are located inside the active site of the X-ray structures 1m2z, 1nhz, 1p93. Figures 6.3(b), 6.3(c) and 6.3(d) show the same side chains in three different IRECS models with different rotamer densities. The reason for the bad docking performance of the IRECS model with $\rho_{rot} = 1$ is that the side chain of Arg611 does not point towards the ligands



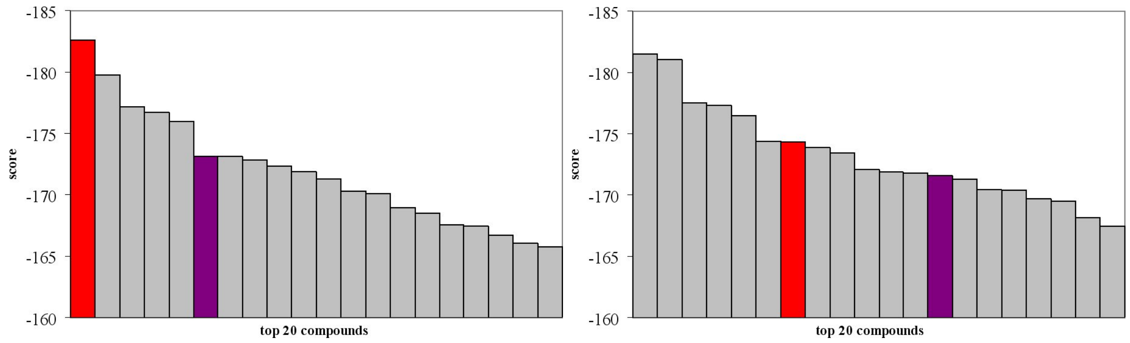
(a) X-ray structures: 1m2z (grey blue), 1nhz (cyan), 1p93 (green)

(b) IRECS model, $\rho_{rot} = 1$

(c) IRECS model, $\rho_{rot} = 2$

(d) IRECS model, $\rho_{rot} = 3$

Figure 6.3: Side chains in the active site of the glutocorticoid receptor



(a) rotamer density=2 (b) rotamer density=3

Figure 6.4: Ranking of the top 20 compounds after screening of HMGR The ranking on the left hand side was done using an IRECS model with $\rho_{rot} = 2$, the ranking on the right hand side used an IRECS model with $\rho_{rot} = 3$. The two active compounds in this subset are colored red and violet, respectively.

as it does in the X-ray structures. Thus the screened compounds cannot form a hydrogen bond to this side chain which results in a decreased binding affinity of active compounds. The required rotamer is offered for Arg611 in the IRECS model with $\rho_{rot} = 2$ (Figure 6.3(c)) which results in better performance of the screening setup using this model. Also Gln642 receives an alternative rotamer which is appropriate for binding different ligands as known from superposition of X-ray structures. The screening performance even increases when using the IRECS model with $\rho_{rot} = 3$ (Figure 6.3(d)) since it contains an alternative rotamer for Met646 which turns away from the center of the active site. This enlarges the volume of the active site and allows the docking of larger ligands.

The screening performance may decrease if too much protein flexibility is allowed during docking. The screening experiments on the protein HMGR are good examples for a decrease in screening performance through docking errors arising from the protein being too flexible during docking. Figure 6.4 shows the scores of the top 1% compounds of the databases that were ranked by docking all compounds into IRECS models with $\rho_{rot} = 2$ and $\rho_{rot} = 3$. The scores of true binders are highlighted in red and violet, respectively. Here, one can see two effects that lead to a lower enrichment for higher flexibility: first, the decoy compounds receive higher scores overall. This is caused by the described higher acceptance rate of all screened compounds and the resulting lowered ability in identifying the true binders. The second effect is a reduced score of the true binders. This effect can be explained by the increasing difficulty for FlexE to find the true binding mode of a ligand with increasing conformational space of the protein. This example illustrates that if all important side-chain conformations are already available in a IRECS model, it is not useful for docking to extend the rotamer ensembles to higher density.

The benefits of side-chain flexibility for screening become apparent if the performances of the screening experiments using rigid and flexible IRECS models are compared. When using rigid IRECS models the enrichment factors reached are clearly lower than those enrichment factors achieved with using flexible IRECS models. This indicates that the softness of the

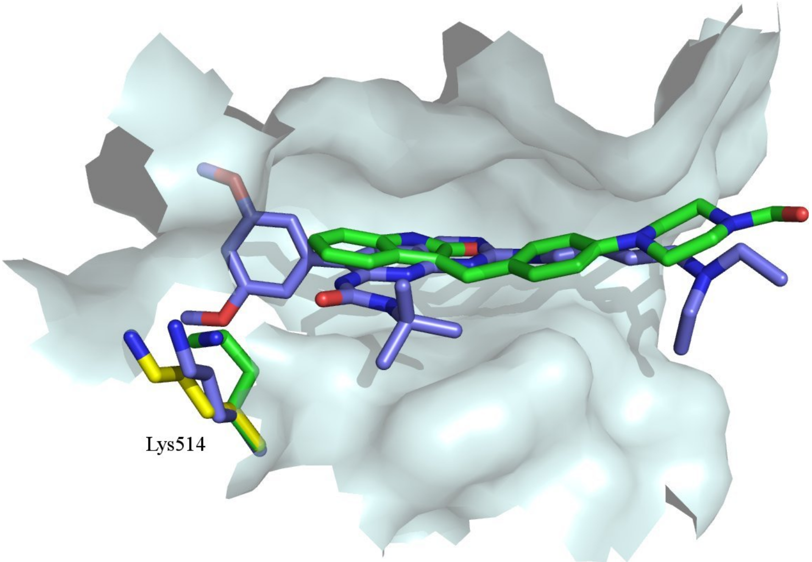


Figure 6.5: Ligands in the active site of FGFr1 interacting with the flexible side chain Lys-514. The ligand of 1agw (SU4984) is colored green, the ligand of 2fgi (PD173074) is colored blue.

ROTA potentials alone is insufficient for handling all required adaptive motions of protein atoms upon ligand binding. In such cases where an inaccurate protein model is used the additional inclusion of side-chain flexibility is necessary for successful screening.

The enrichment factors listed in Table 6.8 show that the performances of the screening setups using flexible IRECS models are quite similar to those of screening setups that use X-ray structures. The lower performance of screening setups that use rigid IRECS indicates that the flexible handling of side chains is essential for this achievement. As the accuracy of IRECS is comparable to that of other side-chain prediction programs (see Table 4.1) it can be assumed that also other side-chain prediction programs using the BBDep (or similar rotamer libraries) are not able to generate rigid protein models that would be suited better for screening with FlexE than the flexible IRECS models are.

On average flexible IRECS models are better suited for screening than rigid IRECS models and ROTA is better suited for scoring complexes with IRECS models, the screening results of the protein FGFr1 shown in Figure A.5 prove that both of these statement need not hold for any individual screening problem. On this protein, the screening setup using F-Score and the rigid IRECS model performs best, although this model has an average $\chi_{1,2}$ accuracy of only 73.3%. In general, this setup performs worst on average both in redocking and virtual screening experiments. On this target, F-Score has a much better performance than ROTA, as the comparison of the enrichment curves (red and cyan curves in Figure A.5) for the X-ray structures and rigid IRECS models using these two scoring functions shows. The advantage of the rigid IRECS model over the X-ray structure in screening can be presumed to occur from a certain conformation of a single side chain that was chosen by IRECS in the rigid case and that deviates from the conformation found in the X-ray structure. Figure 6.5 depicts the active site of FGFr1 as defined in the X-ray structures

1agw (green, from DUD), 2fgi (blue, superposed on 1agw) and the rigid IRECS model (yellow). The figure visualizes the surface of the active site of FGFr1 together with the side chain of Lys514 (left bottom) and the inhibitors SU4984 (from 1agw) and PD173074 (from 2fgi). PD173074 is positioned more deeply inside the active site than SU4984 which causes Lys514 to rotate its side chain and enlarge the cavity for PD173074. This conformation is also present in the IRECS model of FGFr1, although this conformation was optimized using the conformation of SU4984. Although the prediction of IRECS for Lys514 must be considered as wrong in the context of SU4984, the chosen conformation proved to be advantageous for the screening experiments since it models the induced-fit effect observed at Lys514 upon binding compounds like PD173074 that extend deeper into the active site than SU4984.

6.2.3 Runtime

The runtime of a screening experiment is a crucial issue because it is usually desirable to screen as many compounds as possible. The previous sections showed the advantages of inclusion of the protein flexibility in the modeling and docking procedure for redocking and screening performance. However, this increased performance has the price of an extended runtime of each docking attempt for all docked compounds. Table 6.9 shows the setups (program, model and scoring function), the accumulated runtime on a single CPU (Opteron V20z) and the per-ligand runtime of all performed screening runs. Altogether, the DUD target-specific databases contained 117,189 compounds. The 'per-ligand' runtime was calculated by dividing the accumulated runtime of each screening run through this number. The comparison of the runtimes of the different screening setups first reveals that using ROTA instead of F-Score for scoring roughly doubles the required runtime. The extra time originates from the fact that all atoms within a sphere of 10 Å radius around a ligand atom contribute to the score of this atom, so the computational costs of ROTA are higher than those of F-Score.

The second result of the comparison is that FlexE requires much more time for docking than FlexX does. The reason for this is the extra computational time required to sample the conformational space of side chains and ligands simultaneously. This time is mainly consumed by the placement algorithm of FlexX for base fragments of ligands [173] (see Section 2.4.1), which requires a time of $O(n^2)$, with n being the number of interaction points defined in the active site. As each additional conformation of a hydrogen bond donor or acceptor also adds a new interaction geometry to the active site, the number of interaction points into the active site grows linearly with the average number of rotamers that are used for docking.

Table 6.8: Number of targets reaching different enrichment factors for multiple screening setups

(a) considering the top 20% of the ranked libraries

	Model	Rotamer	Scoring	$EF_{20} >$					
Program	source	density	function	1	2	4	8	16	32
FlexX	X-ray	1	F-Score	31	18	5	0	0	0
FlexX	X-ray	1	ROTA	27	12	2	0	0	0
FlexX	IRECS	1	F-Score	24	12	1	0	0	0
FlexX	IRECS	1	ROTA	26	14	1	0	0	0
FlexE	IRECS	2	ROTA	33	17	0	0	0	0
FlexE	IRECS	3	ROTA	31	19	1	0	0	0

(b) considering the top 1% of the ranked libraries

	Model	Rotamer	Scoring	$EF_{1} >$					
Program	source	density	function	1	2	4	8	16	32
FlexX	X-ray	1	F-Score	27	26	18	10	5	0
FlexX	X-ray	1	ROTA	28	25	19	13	3	0
FlexX	IRECS	1	F-Score	18	16	13	6	1	0
FlexX	IRECS	1	ROTA	24	23	15	8	1	0
FlexE	IRECS	2	ROTA	29	26	23	11	2	0
FlexE	IRECS	3	ROTA	29	27	20	11	2	0

(c) considering any subset with highest enrichment factor

	Model	Rotamer	Scoring	$EF_{max} >$					
Program	source	density	function	1	2	4	8	16	32
FlexX	X-ray	1	F-Score	38	31	30	20	17	13
FlexX	X-ray	1	ROTA	40	32	26	22	16	12
FlexX	IRECS	1	F-Score	38	26	16	15	12	9
FlexX	IRECS	1	ROTA	40	31	24	15	10	9
FlexE	IRECS	2	ROTA	40	34	28	24	20	13
FlexE	IRECS	3	ROTA	40	35	26	21	18	15

Table 6.9: Runtimes of screening experiments

	Runtime (single processor)	
Screening setup	Accumulated	Per ligand
FlexX, X-ray, F-Score	21.2 days	15.6 s
FlexX, X-ray, ROTA	47.8 days	35.2 s
FlexX, IRECS-1, F-Score	20.5 days	15.1 s
FlexX, IRECS-1, ROTA	46.1 days	34.0 s
FlexE, IRECS-2, ROTA	228.8 days	168.7 s
FlexE, IRECS-3, ROTA	253.5 days	187.0 s

Chapter 7

Discussion

This chapter first discusses the relevance and applicability of the modeling and docking pipeline (short *pipeline* here) as well as the associated methods in different application scenarios that are likely to occur in structure-based drug design. This discussion also comprises a comparison of the pipeline with existing methods that are applied for same purpose. The validity of some of the key assumptions made throughout this work is checked in the context of these application scenarios. Finally, possible extensions of the pipeline are described with an estimate of their implementation effort and their impact on speed and accuracy of the pipeline. This discussion concentrates on general aspects of the previously presented methods and results, as those topics that are closely related to implementations of the base techniques of the pipeline were already discussed in the respective previous chapters.

Application Scenarios

A researcher planning a screening experiment is usually interested in finding the optimal screening setup given a particular target protein, an initial model of the protein, a screening library and limited resources (manpower, computers and time). This work suggests scanning the proteins of the DUD for a protein most similar to the protein of interest, considering flexibility, hydrophobicity, buriedness of the ligand and overall size. The performance of the pipeline can then be quickly estimated with the enrichment plot for the respective protein, where the quality of the protein model must also be taken into consideration. If such an analysis is not feasible (e.g. because of bad model quality), a more general analysis of the concrete application scenario that takes only protein flexibility and model quality into account can yield valuable insights to the applicability of the modeling and docking pipeline.

There are two primary application scenarios in which the use of the pipeline or its components are likely to perform better than existing methods, as far as suggested by the results that were obtained in this work. Both scenarios have in common that upon scoring of a certain conformation of the protein-ligand complex the position of some of the relevant atoms is – to varying degree – uncertain. The results from the previous chapter also imply that there exists a scenario in which the pipeline is likely to perform worse than other methods. This is the case when (i) a high-quality model of the active site of a protein is available and (ii) the protein is known to be rigid and (iii) a dense sampling mechanism for ligand conformations is applied, as in these cases any modifications by IRECS to the protein side chains can just worsen the structure quality and the softness of ROTA is not of value.

A first scenario that can be characterized with the term 'low quality protein model' comprises many of those studies in which a protein model is derived by either homology modeling, averaging of NMR models or low resolution X-ray crystallography. This scenario requires a good quality of the backbone conformation, whereas some of the side chains are tolerated to have wrong conformations. This is a quite frequent scenario for screening setups that cannot be executed using a high resolution X-ray structure, e.g. [39, 50, 99, 106, 138, 152, 213, 212, 229]. Some of these studies apply docking algorithms that simulate protein flexibility or apply softer scoring functions, but many of these studies just ignore the inaccuracy of the protein model during docking. This renders their docking setup quite similar to the setup proposed in this work which uses just FlexX, F-Score and rigid protein models. This setup performed worst in most redocking and screening experiments undertaken here which displays the inferiority of this setup compared to the other setups proposed in this work. It can also be assumed that a number of docking studies that failed because of wrongly chosen docking setups were not published in the past. The pipeline now offers a combined procedure for dealing with this issue: a remodeling of the side chains with IRECS and predicting two conformations per side chain on average has a good chance of providing correct conformations for side chains, from which FlexE can select the best matching ones upon docking. The softness of ROTA can account for the small errors occurring from the slightly inaccurate backbone conformation and the discrete rotamer assumption by using rotamers from the BBDep. Another option would be to apply the MOBILE (see Section 2.4.3) or IFD (see Section 2.4.3) approaches and improve the protein conformation with the help of a known ligand. Although these techniques have the drawback of imprinting the binding mode and topology of this ligand onto the active site they have the potential of improving the later docking performance significantly. Both methods have the caveat that if the initial docking of the known inhibitor results in a wrong conformation, the subsequent modeling procedure is misdirected and will surely create a wrong protein model. The probability of this is substantial since MOBILE uses ensembles of homology models generated with MODELLER that often provide wrong side-chain conformations [216] and usually lack alternative conformations. IFD is simply disregarding the part of all side chains beyond their C_β atom for the initial placement (effectively mutating all residues to alanines except glycines). The pipeline can support both MOBILE and IFD in that it allows for docking the known inhibitor with higher accuracy than it would be possible using the techniques that were initially proposed. Finally, the determining factor of the applicability of the pipeline remains the expected amount of coordinate uncertainty of atoms: if only small positional deviations of protein atoms are expected ROTA can handle them. The application of the complete pipeline is therefore advised if wrong rotamers are expected and a near-native backbone conformation is available.

A second scenario can be characterized with the term 'induced fit'. In this scenario the protein conformation is not adapted to the conformation of the ligand. A ligand docking is likely to fail if the protein is either in the apo conformation or a holo conformation that is adapted to another ligand. This is a well-known scenario for which docking programs were developed that can account for protein flexibility during docking (see Section 2.4.2). Ensembles of preset protein conformations for sampling the protein conformational space are both used by docking programs e.g. FlexE or different approaches using DOCK [88, 109, 134, 225]) and approaches that implement the wide-spread serial docking strategy (see Section 2.4.2). IRECS can support these approaches for the frequent cases where there are no multiple X-ray structures available in generating ensembles of protein conformations with alternative side-chain conformations. As previously stated (see Section 4.1) IRECS is optimized for achieving a good trade off between coverage of the conformational space and the number of required samples, a feature that is most important for time-critical applications. Alternative computational methods that are applied for sampling the conformational space of the protein (Monte Carlo sampling, snapshots from molecular dynamics simulations) often require too many samples to cover the conformational space therefore are less suitable for directly providing input for ensemble docking approaches. Whereas the rotamers of IRECS can be adapted to a ligand conformation in terms of selecting the best fitting rotamers from BBDep, the individual rotamers in BBDep are not adapted to a certain environment and therefore the IRECS ensembles are usually less well adapted to a ligand than a similar ensemble drawn from X-ray structures. This issue can be handled by running energy minimizations on the IRECS models or by applying a soft scoring function like ROTA during docking as it was previously shown by Ferrari and Shoichet [51]. Sometimes it can also be helpful to use a protein conformation for docking that is not adapted towards a certain class of ligands. Actually, molecular similarity method like Feature Trees [170] or MolPrint [8, 9] are often better suited for screening for active compounds that are similar to a known ligand than docking techniques (Andreas Steffen, personal communication). Using the unadapted rotamers of the IRECS models prevent the screening experiment from becoming biased towards compounds that are too similar to the co-crystallized ligand and therefore increases the chance of finding new classes of compound scaffolds, which recently has become increasingly important in pharmaceutical research. Again, the application of IRECS and ROTA to a single backbone is only advised if major conformational changes of the backbone are not expected. Otherwise, one could sample multiple backbone conformations (e.g. with MODELLER) and then let IRECS sample the relevant side-chain conformations.

Consequences of Ignoring Backbone Flexibility

The restriction of ROTA and IRECS to side-chain flexibility ignoring backbone flexibility is also their strongest limitation. If the induced-fit effects also require that the backbone changes its conformation before the ligand can bind, then this pipeline is quite likely to fail. The benchmarking results of Wallner and Elofsson [216] justify the separate treatment of side-chain and backbone flexibility during homology modeling. It has been shown that programs specializing in side-chain prediction can improve protein models that are generated by programs that treating the whole protein as flexible (like those presented in Section 2.3.1). The decision for this limitation was made quite in the early stages of the project, as this limitation was a requirement that was both important for (i) the comprehensive sampling of modeling failures that were necessary for the derivation of ROTA and (ii) for efficient optimization with IRECS. Furthermore, it was previously reported that FlexE achieves better accuracy if protein flexibility is limited to rotations of side chains and small loop changes [162, 201].

A quite straightforward extension of the derivation procedure of ROTA would use MODELLER for generating decoy conformations of the protein backbone before decoy conformations for side chains are generated. Then ROTA would also be able to score backbone-backbone interactions. An extension of IRECS would allow for multiple preconfigured backbone conformations as input and define additional ensembles of conformations for backbone segments, as it was done by the FlexE docking program. Such an extension however would slow down IRECS and largely increase the number of atom positions of the generated model, which in turn would require either exchanging the placement algorithm implemented in FlexE or the application of another docking program.

Consequences of Overlapping Training and Test Sets of Protein X-ray Structures

The significance of evaluation studies is always and rightfully questioned if the same or highly correlated data are used for both training and testing purpose [74]. Since the number of high-quality protein structures is quite limited on one side and the number of such structures required for training the different scoring functions and algorithms of this work is high on the other hand, not all such overlaps could be prevented. Although this issue is usually ignored in the field whenever both training and tests sets were not derived in the context of the same study (see previous evaluations of scoring functions presented in Section 3.7), an attempt is made here to name the most critical overlaps and estimate the effects of these overlaps on the evaluation results.

Throughout this work, eight different sets of X-ray protein structures were used for either training (see Sections 3.1, 3.2, 4.2.2, 5.3) or testing (see Sections 3.7, 4.5, 4.6, 6.1.1) purposes. Further sets of structures were used by Dunbrack and Cohen to derive the BBDep [43], and other sets of protein structures were used to parameterize FlexX and train F-Score. All of these sets were created with the goal of capturing a representative subset of the whole known structural space of proteins. As usually, structures with highest accuracy were chosen for the sets. There exists a number of direct overlaps between training and test sets, e.g. the Top-500 set has 31 overlaps with the test set in Section 4.5 and the training set of the BBDep has 83 overlaps with the same test set (with 160 structures) and 146 overlaps with the test set used in Section 4.6 (with 447 structures). Related protein structures from different sets are often not exactly the same (same PDB ID), but are structurally similar (e.g. through homology relationship) which results in a much higher number of indirect (and unidentified) overlaps. Although some attempts were performed to circumvent direct overlaps – the 10-fold cross validation in Section 5.5, the splitting of test and training sets in Section 4.5 – no attempts were undertaken to prevent indirect overlaps.

There exist two critical overlaps: (i) the overlaps of training and test sets of ROTA (DUD [85] and test set of Wang et al. [223]) and (ii) the overlaps between the training set of the BBDep and the test set for IRECS. The influence of a single overlap depends both on the size of training set and its particular influence on the prediction of a single evaluated feature. In the case of ROTA, the energy score is a combination of many (usually more than 100) individual predictions based on the ROTA potentials. The effect that a single structure has on the shape of these potentials is strongest on potentials for rare atom type combinations (e.g. ligand halogen with an atom of Tryptophan) at close distances. These signals are so rare that a signal from overlapping structures has a strong influence on the contribution of this potential to the ROTA score. However, the contribution of individual potentials to the overall ROTA score remains low and it can be estimated that this holds also true for overlap effects, as the percentage of near-atom contacts is quite low considering the 10.0 Å distance cutoff of ROTA. The situation is quite different for the overlaps observed for the training set of the BBDep and the IRECS test set. Here, the intense fragmentation of the feature space (see Section 2.3.2) enables a single side chain to exert a strong influence on the rotamer population that is predicted for a sparsely populated area of the Ramachandran plot. Since these areas mostly correspond to protein substructures that are much harder to predict than helices or sheets, IRECS and SCWRL can greatly benefit from any overlap that occurs between side chains in such areas. This implies that the results that were presented in Figure 4.1 are somewhat biased towards IRECS and SCWRL and disfavor SCAP. Considering the small differences in performance shown in this figure, no well-founded suggestion for a single side-chain prediction program can be given for predicting rigid protein structures. As the CHARMM force field used by SCAP was developed over years by a whole community of researches that also were able to fit parameters empirically to various structural observations, it appears to be infeasible to setup a fair comparison using any structural data that were previously published.

Future Extensions

As ROTA has shown superior performance in estimating binding affinity of protein-ligand complexes that were not optimized for scoring (see Section 3.7.2), ROTA is especially helpful in situations in which no elaborate optimization of the complex conformation can be afforded. One such application would be the screening for peptides binding to the major histocompatibility complex (MHC). Side chains of each query peptide sequence can be mutated and optimized with IRECS in concert with the side chains of the MHC. A fast IRECS optimization can generate a preliminary complex structure that then can be scored using ROTA without requiring further optimization. Since peptides bound to MHCs usually exhibit only few variation of their backbone conformation, such a system can be realized by implementing an extension of IRECS that would allow for considering a small set of backbone conformations serially.

IRECS is able to determine the individual flexibility of side chains, a feature that was exploited in this work to appropriately size rotamer ensembles for docking. This ability can also be used for general flexibility analysis that concentrates on side-chain flexibility. Such information could be used in various applications: if the flexibility of side chains is computed before and after docking, the loss of entropy on the protein side upon complex building can be estimated. This can be used as an additional component to scoring functions and thereby increasing prediction accuracy of these functions. Also, any measure of side-chain flexibility can be used as confidence measure for structural properties (atom positions, surface descriptors, interaction scores) that were calculated using a single, rigid protein structure.

Chapter 8

Conclusion

A number of techniques were presented in this work that can support docking into homology models and into flexible proteins. These techniques can be combined as shown in Chapter 6 to a complete modeling and docking pipeline or be applied as stand-alone applications. The scoring function ROTA (see Chapter 3) plays a central role in this concept. A special derivation technique was developed for ROTA so that ROTA is able to tolerate minor failures of atom positions and is especially sensitive to false rotamer states and false ligand placements. ROTA can maintain a high accuracy in predicting binding affinity of protein-ligand complexes in cases in which the protein and ligand conformation are not adapted to each other. This ability is especially useful for modeling induced-fit effects, which was confirmed by using ROTA as a guiding scoring function for docking with FlexE in Chapter 6.

One central prerequisite for flexible docking with FlexE is a predefined ensemble of rotamers for flexible side chains. However, the runtime of FlexE and the false-positive rate in a screening experiment would increase if such ensembles contained more rotamers than the flexibility of their respective side chain can justify. Therefore a probabilistic model of side-chain flexibility was developed and the problem of modeling side-chain flexibility with a limited set of rotamers was formulated in Chapter 4. The side-chain prediction algorithm IRECS was developed, which is able to generate approximate solutions for this problem. IRECS can accurately predict side-chain conformations for both rigid and flexible protein models and reduce rotamer ensembles so that the resulting ensembles represent the conformational space of side chains. This renders IRECS a suitable preprocessing tool for FlexE for such protein targets that do not exhibit strong backbone flexibility in the active site but have flexible side chains or for proteins for which no protein models with highly accurate side-chain conformations could be generated.

Side-chain prediction becomes especially hard in cases in which the original backbone conformation is not present in the model but instead a conformation from a homologous protein is used as a template for modeling. The rotamer-lock algorithm was designed for supporting IRECS in such cases and is described in Chapter 5. This algorithm allows for incorporating the structural information available for the template side chains in the IRECS optimization procedure for the target side chains. However, the overall contribution of the rotamer-lock algorithm to the prediction accuracy of IRECS is low.

Like for other current side-chain prediction tools, the accuracy of IRECS in modeling rigid protein structures is not high enough for docking purposes, since about one out of six predicted side-chain conformation has a wrong orientation in a rigid model. Since such errors can greatly spoil a docking attempt, a number of additional rotamers is predicted so that the overall chance is increased that each ensemble of rotamers contains at least one rotamer with a correct orientation. ROTA, IRECS and FlexE are supposed to act in concert in a single docking pipeline, and thus an extensive evaluation was carried out that allowed for inspecting the influence of ensemble sizes on docking accuracy. The docking program FlexE was extended in Chapter 6 so that it is able to use IRECS models for simulating protein flexibility and quickly scoring protein-ligand interactions with ROTA. An extract of the screening database DUD was prepared so that IRECS models and FlexE docking runs could be executed automatically, allowing to test a substantial number of modeling and screening setups. The major outcome of this evaluation was that two rotamers per side-chain on average are sufficient to guarantee a high performance of FlexE. Also, the advantages of using ROTA and IRECS models could be shown for cases in which flexible side chains are part of the active site or in which side chains in rigid protein models lack accurate conformations. The results also enabled the identification of application scenarios in which the pipeline is most likely to achieve a higher accuracy than traditional modeling and docking approaches or where it can at least support other established approaches.

Bibliography

[1] R. Abagyan, M. Totrov, and D. Kutznetsov. ICM - A new method for protein modelling and design: Applications to docking and structure predication from the distorted native conformation. *J Comput Chem*, 15(5):488–506, 1994.

[2] E. E. Abola, A Balroch, W.C. Barker, S. Beck, H. Benson, D.A. Berman, G. Cameron, C. Cantor, S. Doubet, T. J. P. Hubbard, T.A. Jones, G. J. Kleywegt, A. S. Kolaskar, A. Van Kuik, A. M. Lesk, H.-W. Mewes, D. Neuhaus, F. Pfeiffer, L. F. TenEyck, R.J. Simpson, G. Stoesser, J. L. Sussman, Y. Tateno, A. Tsugita, E. L. Ulrich, and J. F. G. Viiegenthart. Quality control in databanks for molecular biology. *Bioessays*, 22(11):1024–1034, 2000.

[3] F. H. Allen. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Cryst B*, 58(Pt 3 Pt 1):380–388, 2002.

[4] E. Althaus, O. Kohlbacher, H.P. Lenhof, and P. Müller. A combinatorial approach to protein docking with flexible side chains. *J Comput Biol*, 9(4), 2002.

[5] A. C. Anderson, R. H. O'Neil, T. S. Surti, and R. M. Stroud. Approaches to solving the rigid receptor problem by identifying a minimal set of flexible residues during ligand docking. *Chem Biol*, 8(5):445–457, 2001.

[6] K. C. D. Bahadur, E.Tomita, J. Suzuki, and T. Akutsu. Protein side-chain packing problem: a maximum edge-weight clique algorithmic approach. *J Bioinform Comput Biol*, 3(1):103–126, 2005.

[7] D. Baker and A. Sali. Protein structure prediction and structural genomics. *Science*, 294(5540):93–96, 2001.

[8] A. Bender, H.Y. Mussa, R.C. Glen, and S. Reiling. Molecular similarity searching using atom environments, information-based feature selection, and a naive bayesian classifier. *J Chem Inf Comput Sci*, 44:170–178, 2004.

[9] A. Bender, H.Y. Mussa, R.C. Glen, and S. Reiling. Similarity searching of chemical databases using atom environment descriptors: evaluation of performance. *J Chem Inf Comput Sci*, 44:1708–1718, 2004.

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

[11] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. The Protein Data Bank. A computer-based archival file for macromolecular structures. *Eur J Biochem*, 80(2):319–324, 1977.

[12] H. J. Böhm. The computer program LUDI: a new method for the de novo design of enzyme inhibitors. *J Comput Aided Mol Des*, 6(1):61–78, 1992.

[13] H. J. Böhm. LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads. *J Comput Aided Mol Des*, 6(6):593–606, 1992.

[14] H. J. Böhm. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J Comput Aided Mol Des*, 8(3):243–256, 1994.

[15] H. J. Böhm. Prediction of binding constants of protein ligands: A fast method for the prioritization of hits obtained from de novo design or 3d database search programs. *Journal of Computer-Aided Molecular Design*, 12:309–309, 1998.

[16] H. J. Böhm, G. Klebe, and H. Kubinyi. *Wirkstoffdesign*. Spektrum Akademischer Verlag GmbH, 1996.

[17] E. Bindewald and J. Skolnick. A scoring function for docking ligands to low-resolution protein structures. *J Comput Chem*, 26(4):374–383, 2005.

[18] J. M. Blaney and J. S. Dixon. *DockIt, version 1.0; Metaphorics, LLC: Mission Viejo, CA, www.metaphorics.com/products/dockit.html.*

[19] M. J. Bonifácio, M. Archer, M. L. Rodrigues, P. M. Matias, D. A. Learmonth, M. A. Carrondo, and P. Soares-Da-Silva. Kinetics and crystal structure of catechol-o-methyltransferase complex with co-substrate and a novel inhibitor with potential therapeutic application. *Mol Pharmacol*, 62(4):795–805, 2002.

[20] P. E. Bourne and H. Weissig. *Structural Bioinformatics*. Wiley-Liss, Inc., 2003.

[21] M. J. Bower, F. E. Cohen, and R. L. Dunbrack. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. *J Mol Biol*, 267(5):1268–1282, 1997.

[22] C. Brandon and J. Tooze. *Introduction to Protein Structure*. Garland Publishing, Inc., 1999.

[23] C. Bron and J. Kerbosch. Algorithm 457: finding all cliques of an undirected graph. *CACM*, 16:575–577, 1973.

[24] N. Brooijmans and I. D. Kuntz. Molecular recognition and docking algorithms. *Annu Rev Biophys Biomol Struct*, 32:335–373, 2003.

[25] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J Comp Chem*, 4:187–217, 1983.

[26] A. A. Canutescu, A. A. Shelenkov, and R. L. Dunbrack. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci*, 12(9):2001–2014, 2003.

[27] H. A. Carlson. Protein flexibility and drug design: how to hit a moving target. *Curr Opin Chem Biol*, 6(4):447–452, 2002.

[28] H. A. Carlson. Protein flexibility is an important component of structure-based drug discovery. *Curr Pharm Des*, 8(17):1571–1578, 2002.

[29] O. Carugo and P. Argos. Protein-protein crystal-packing contacts. *Protein Sci*, 6(10):2261–2263, 1997.

[30] C. N. Cavasotto and R. A. Abagyan. Protein flexibility in ligand docking and virtual screening to protein kinases. *J Mol Biol*, 337(1):209–225, 2004.

[31] V. Cerny. A thermodynamical approach to the travelling salesman problem: an efficient simulation algorithm. *J Optimiz Theory App*, 45:41–51, 1985.

[32] Chemical Computing Group, Montreal, Canada. http://www.chemcomp.com/. *Standard docking routine implemented in MOE, version 2002.03.*

[33] H. Claußen, C. Buning, M. Rarey, and T. Lengauer. FlexE: efficient molecular docking considering protein structure variations. *J Mol Biol*, 308(2):377–395, 2001.

[34] A. M. Davis, S. J. Teague, and G. J. Kleywegt. Application and limitations of x-ray crystallographic data in structure-based ligand and drug design. *Angew Chem Int Ed Engl*, 42(24):2718–2736, 2003.

[35] WL DeLano. The pymol molecular graphics system. *DeLano Scientific, San Carlos, CA, USA*, 2002.

[36] J. Desmet, M. Demaeyer, B. Hazes, and I. Lasters. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*, 356:539–542, 1992.

[37] J. Desmet, J. Spriet, and I. Lasters. Fast and accurate side-chain topology and energy refinement (faster) as a new method for protein structure optimization. *Proteins*, 48(1):31–43, Jul 2002.

[38] C. DeWeese-Scott and J. Moult. Molecular modeling of protein function regions. *Proteins*, 55(4):942–961, 2004.

[39] D. J. Diller and R. X. Li. Kinases, homology models, and high throughput docking. *J Med Chem*, 46(22):4638–4647, 2003.

[40] F. S. Domingues, J. Rahnenführer, and T. Lengauer. Automated clustering of ensembles of alternative models in protein structure databases. *Protein Eng Des Sel*, 17:537–43, 2004.

[41] R. L. Dunbrack. Rotamer libraries in the 21st century. *Curr Opin Struct Biol*, 12(4):431–440, 2002.

[42] R. L. Dunbrack. *Homology Modeling in Biology and Medicine. In: Bioinformatics - From Genome to Therapies, Volume 2, Ed.: T. Lengauer.* Wiley-VCH Verlag, 2007.

[43] R. L. Dunbrack and F. E. Cohen. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci*, 6(8):1661–1681, 1997.

[44] M. D. Eldridge, C. W. Murray, T. R. Auton, G. V. Paolini, and R. P. Mee. Empirical scoring functions: I. the development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J Comput Aided Mol Des*, 11(5):425–445, 1997.

[45] O. Eriksson, Y. Zhou, and A. Elofsson. Side chain-positioning as an integer programming problem. *WABI 01: Proceedings of the First International Workshop on Algorithms in Bioinformatics*, pages 128–141, 2001.

[46] A. Evers, H. Gohlke, and G. Klebe. Ligand-supported homology modelling of protein binding-sites using knowledge-based potentials. *J Mol Biol*, 334(2):327–345, 2003.

[47] A. Evers and G. Klebe. Ligand-supported homology modeling of g-protein-coupled receptor sites: Models sufficient for successful virtual screening. *Angew Chem Int Ed Engl*, 43(2):248–251, 2004.

[48] A. Evers and G. Klebe. Successful virtual screening for a submicromolar antagonist of the neurokinin-1 receptor based on a ligand-supported homology model. *J Med Chem*, 47(22):5381–5392, 2004.

[49] T. J. A. Ewing, S. Makino, A. G. Skillman, and I. D. Kuntz. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J Comput Aid Mol Des*, 15:411–428, 2001.

[50] M. X. Fernandes, V. Kairys, and M. K. Gilson. Comparing ligand interactions with multiple receptors via serial docking. *Journal of Chemical Information and Computer Sciences*, 44(6):1961–1970, 2004.

[51] A. M. Ferrari, B. Q. Wei, L. Costantino, and B. K. Shoichet. Soft docking and multiple receptor conformations in virtual screening. *J Med Chem*, 47(21):5076–5084, 2004.

[52] E. Fischer. Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte der Deutschen Chemischen Gesellschaft*, 27:2985–2993, 1894.

[53] M. A. Fox and J. K. Whitesell. *Organische Chemie*. Spektrum Akademischer Verlag, 1995.

[54] Y. Freund and L. Mason. The alternating decision tree learning algorithm. In *Proceeding of the Sixteenth International Conference on Machine Learning, Bled, Slovenia: 124-133*, 1999.

[55] R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry, D. E. Shaw, P. Francis, and P. S. Shenkin. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem*, 47(7):1739–1749, 2004.

[56] T. M. Frimurer, G. H. Peters, L. F. Iversen, H. S. Andersen, N. Peter H. Møller, and O. H. Olsen. Ligand-induced conformational changes: improved predictions of ligand binding conformations and affinities. *Biophys J*, 84(4):2273–2281, 2003.

[57] B. R. Gaines and P. Compton. Induction of ripple-down rules applied to modeling large databases. *J Intell Inf Syst*, 5:211–228, 1995.

[58] E. Gallicchio, L. Y. Zhang, and R. M. Levy. The SGB/NP hydration free energy model based on the surface generalized born solvent reaction field and novel nonpolar hydration free energy estimators. *J Comp Chemp*, 23(5):517–529, 2002.

[59] J. Gasteiger, C. Rudolph, and J. Sadowski. Automatic generation of 3D-atomic coordinates for organic molecules. *Tetrahedron Comput Method*, 3:537–547, 1992.

[60] D. K. Gehlhaar, G. M. Verkhivker, P. A. Rejto, C. J. Sherman, D. B. Fogel, L. J. Fogel, and S. T. Freer. Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. *Chem Biol*, 2(5):317–324, 1995.

[61] A. Ghosh, C. S. Rapp, and R. A. Friesner. Generalized born model based on a surface integral formulation. *J Phys Chem B*, 102(52):10983–10990, 1998.

[62] M. K. Gilson and H.-X. Zhou. Calculation of protein-ligand binding affinities. *Annu Rev Bioph Biom*, 36:21–42, 2007.

[63] H. Gohlke, M. Hendlich, and G. Klebe. Knowledge-based scoring function to predict protein-ligand interactions. *J Mol Biol*, 295(2):337–356, 2000.

[64] R. F. Goldstein. Efficient rotamer elimination applied to protein side-chains and related spin glasses. *Biophys J*, 66(5):1335–1340, 1994.

[65] D. Benjamin Gordon, Geoffrey K. Hom, Stephen L. Mayo, and Niles A. Pierce. Exact rotamer optimization for protein design. *J Comput Chem*, 24(2):232–243, 2003.

[66] J. J. Gray, S. E. Moughan, C. Wang, O. Schueler-Furman, B. Kuhlman, C. A. Rohl, and D. Baker. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J Mol Biol*, 331:281 – 299, 2003.

[67] S. Grüneberg, B. Wendt, and G. Klebe. Subnanomolar inhibitors from computer screening: A model study using human Carbonic Anhydrase II. *Angew Chem Int Ed Engl*, 40(2):389–393, 2001.

[68] T. A. Halgren. Merck molecular force field. I. basis, form, scope, parameterization, and performance of MMFF94. *J Comp Chem*, 17:490–519, 1998.

[69] T. A. Halgren, R. B. Murphy, R. A. Friesner, H. S. Beard, L. L. Frye, W. T. Pollard, and J. L. Banks. Glide: a new approach for rapid, accurate docking and scoring. 2. enrichment factors in database screening. *J Med Chem*, 47(7):1750–1759, 2004.

[70] I. Halperin, B. Ma, H. Wolfson, and R. Nussinov. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins*, 47(4):409–443, 2002.

[71] C. Hartmann. Erweiterung des Docking-Programms FlexE durch das Modell der Mean Field-Theorie. Master's thesis, Rheinische Friedrich-Wilhelms-Universität, 2004.

[72] C. Hartmann, I. Antes, and T. Lengauer. IRECS: A new algorithm for the selection of most probable ensembles of side-chain conformations in protein models. *Protein Sci*, 16(7):1294–1307, 2007.

[73] C. Hartmann, I. Antes, and T. Lengauer. IRECS: Prediction of side-chain conformation ensembles. In *Poster at the Drug Discovery Workshop, Marburg*, 2007.

[74] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2001.

[75] M. Hendlich. Databases for protein-ligand complexes. *Acta Cryst*, D54:1178–1182, 1998.

[76] M. Hendlich, A. Bergner, J. Günther, and G. Klebe. Relibase: design and development of a database for comprehensive analysis of protein-ligand interactions. *J Mol Biol*, 326(2):607–620, 2003.

[77] A. Hillisch, L. F. Pineda, and R. Hilgenfeld. Utility of homology models in the drug discovery process. *Drug Discov Today*, 9(15):659–669, 2004.

[78] RV Hogg and AT Craig. *Introduction to Mathematical Statistics, 5th ed.* Prentice Hall, 1995.

[79] L. Holm, C. Ouzounis, C. Sander, G. Tuparev, and G. Vriend. A database of protein structure families with common folding motifs. *Protein Sci*, 1(12):1691–1698, 1992.

[80] L. Holm and C. Sander. The FSSP database: fold classification based on structure-structure alignment of proteins. *Nucleic Acids Res*, 24(1):206–209, 1996.

[81] R. C. Holte. Very simple classification rules perform well on most commonly used datasets. *Mach Learn*, 11:63–91, 1993.

[82] R. W. Hooft, G. Vriend, C. Sander, and E. E. Abola. Errors in protein structures. *Nature*, 381:272–272, 1996.

[83] E. S. Huang, P. Koehl, M. Levitt, R. V. Pappu, and J. W. Ponder. Accuracy of side-chain prediction upon near-native protein backbones generated by ab initio folding methods. *Proteins*, 33(2):204–217, 1998.

[84] N. Huang and J. J. Irwin. personal communication. 2007.

[85] N. Huang, B. K. Shoichet, and J. J. Irwin. Benchmarking sets for molecular docking. *J Med Chem*, 49(23):6789–6801, 2006.

[86] S. Y. Huang and X. Zou. An iterative knowledge-based scoring function to predict protein-ligand interactions: I. Derivation of interaction potentials. *J Comput Chem*, 2006.

[87] S. Y. Huang and X. Zou. An iterative knowledge-based scoring function to predict protein-ligand interactions: II. Validation of the scoring function. *J Comput Chem*, 2006.

[88] S. Y. Huang and X. Zou. Ensemble docking of multiple protein structures: considering protein structural variations in molecular docking. *Proteins*, 66(2):399–421, 2007.

[89] S. J. Hubbard, S. F. Campbell, and J. M. Thornton. Molecular recognition. Conformational analysis of limited proteolytic sites and serine proteinase protein inhibitors. *J Mol Biol*, 220(2):507–530, 1991.

[90] J. J. Irwin and B. K. Shoichet. ZINC–a free database of commercially available compounds for virtual screening. *J Chem Inf Model*, 45(1):177–182, 2005.

[91] M. Jacobson and A. Sali. Comparative protein structure modeling and its applications to drug discovery. *Annu Rep Med Chem*, 39:259–276, 2004.

[92] M. P. Jacobson, R. A. Friesner, Z. Xiang, and B. Honig. On the role of the crystal environment in determining protein side-chain conformations. *J Mol Biol*, 320(3):597–608, 2002.

[93] M. P. Jacobson, G. A. Kaminski, R. A. Friesner, and C. S. Rapp. Force field validation using protein side chain prediction. *J Phys Chem B*, 106(44):11673–11680, 2002.

[94] M. P. Jacobson, D. L. Pincus, C. S. Rapp, T. J. F. Day, B. Honig, D. E. Shaw, and R. A. Friesner. A hierarchical approach to all-atom protein loop prediction. *Proteins*, 55(2):351–367, 2004.

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

[96] G. Jones, P. Willett, and R. C. Glen. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J Mol Biol*, 245(1):43–53, 1995.

[97] G. Jones, P. Willett, R. C. Glen, A. R. Leach, and R. Taylor. Development and validation of a genetic algorithm for flexible docking. *J Mol Biol*, 267(3):727–748, 1997.

[98] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids. *J Am Chem Soc*, 118(45):11225–11236, 1996.

[99] V. Kairys, M. X. Fernandes, and M. K. Gilson. Screening drug-like compounds by docking to homology models: a systematic study. *J Chem Inf Model*, 46(1):365–379, 2006.

[100] G. A. Kaminski, R. A. Friesner, J. Tirado-Rives, and W. L. Jorgensen. Evaluation and reparametrization of the opls-aa force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J Phys Chem B*, 105(28):6474–6487, 2001.

[101] E. Kellenberger, J. Rodrigo, P. Muller, and D. Rognan. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins*, 57(2):225–242, 2004.

[102] D. E. Kim, D. Chivian, and D. Baker. Protein structure prediction and analysis using the robetta server. *Nucleic Acids Res*, 32(Web Server issue):W526–W531, 2004.

[103] C. L. Kingsford, B. Chazelle, and M. Singh. Solving and analyzing side-chain positioning problems using linear and integer programming. *Bioinformatics*, 21(7):1028–1036, 2005.

[104] S. Kirkpatrick, C. D. Jr. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598), 1983.

[105] J. G. Kirkwood. Statistical mechanics of fluid mixture. *J Chem Phys*, 3:300–313, 1935.

[106] R. Kiss, Z. Kovári, and G. M. Keseru. Homology modelling and binding site mapping of the human histamine H1 receptor. *Eur J Med Chem*, 39(11):959–967, 2004.

[107] G. Klebe and T. Mietzner. A fast and efficient method to generate biologically relevant conformations. *J Comput Aided Mol Des*, 8:583–606, 1994.

[108] G. J. Kleywegt. Validation of protein crystal structures. *Acta Cryst*, D56:249–265, 2000.

[109] R. M. Knegtel, I. D. Kuntz, and C. M. Oshiro. Molecular docking to ensembles of protein structures. *J Mol Biol*, 266(2):424–440, 1997.

[110] P. Koehl and M. Delarue. Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J Mol Biol*, 239(2):249–275, 1994.

[111] P. Koehl and M. Delarue. A self consistent mean field approach to simultaneous gap closure and side-chain positioning in homology modelling. *Nat Struct Biol*, 2(2):163–170, 1995.

[112] P. Koehl and M. Delarue. Mean-field minimization methods for biological macromolecules. *Curr Opin Struct Biol*, 6(2):222–226, 1996.

[113] R. Kohavi. The power of decision tables. In *Lecture Notes In Computer Science; Vol. 912; Proceedings of the 8th European Conference on Machine Learning: 174 - 189*, 1995.

[114] I. P. Korndörfer, S. Schlehuber, and A. Skerra. Structural mechanism of specific ligand recognition by a lipocalin tailored for the complexation of digoxigenin. *J Mol Biol*, 330(2):385–396, 2003.

[115] D. E. Koshland. Application of a theory of enzyme specificity to protein synthesis. *Proc Natl Acad Sci USA*, 44:98–104, 1958.

[116] O. Kraemer, I. Hazemann, A. D. Podjarny, and G. Klebe. Virtual screening for inhibitors of human aldose reductase. *Proteins*, 55(4):814–823, 2004.

[117] B. Kramer, M. Rarey, and T. Lengauer. Evaluation of the FlexX incremental construction algorithm for protein-ligand docking. *Proteins*, 37(2):228–241, 1999.

[118] A. Krammer, P. D. Kirchhoff, X. Jiang, C. M. Venkatachalam, and M. Waldman. LigScore: a novel scoring function for predicting binding affinities. *J Mol Graph Model*, 23(5):395–407, 2005.

[119] I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin. A geometric approach to macromolecule-ligand interactions. *J Mol Biol*, 161:269–288, 1982.

[120] M. A. Kurowski and J. M. Bujnicki. GeneSilico protein structure prediction meta-server. *Nucleic Acids Res*, 31(13):3305–3307, 2003.

[121] M. H. Lambert. *Docking Conformationally Flexible Molecules into Protein Binding Sites. In: Practical Application of Computer-Aided Drug Design, Ed.: P. S. Charifson,*. Dekker: New York, 1997.

[122] I. Lasters, M. De Maeyer, and J. Desmet. Enhanced dead-end elimination in the search for the global minimum energy conformation of a collection of protein side chains. *Protein Eng*, 8(8):815–822, 1995.

[123] C. A. Laughton. Prediction of protein side-chain conformations from local three-dimensional homology relationships. *J Mol Biol*, 235(3):1088–1097, 1994.

[124] T. Lazaridis and M. Karplus. Effective energy functions for protein structure prediction. *Proteins*, 35(2):133–152, 1999.

[125] A. R. Leach and A. P. Lemon. Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. *Proteins*, 33(2):227–239, 1998.

[126] B. Lee and F. M. Richards. The interpretation of protein structures: estimation of static accessibility. *J Mol Biol*, 55:379–400, 1971.

[127] J. E. Lennard-Jones. Cohesion. *Proceedings of the Physical Society*, 43:461–482, 1931.

[128] W. Li and A.Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006.

[129] S. Liang and N. V. Grishin. Side-chain modeling with an optimized scoring function. *Protein Sci*, 11(2):322–331, 2002.

[130] C. Lin, A. D. Kwong, and R. B. Perni. Discovery and development of VX-950, a novel, covalent, and reversible inhibitor of hepatitis C virus NS3.4A serine protease. *Infect Disord Drug Targets*, 6(1):3–16, 2006.

[131] C. Lipinski and A. Hopkins. Navigating chemical space for biology and medicine. *Nature*, 432(7019):855–861, 2004.

[132] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Del Rev*, 46:3–26, 2001.

[133] L. L. Looger and H. W. Hellinga. Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: implications for protein design and structural genomics. *J Mol Biol*, 307(1):429–445, 2001.

[134] D.. Lorber, M. K. Udo, and B. K. Shoichet. Protein-protein docking with multiple residue conformations and residue substitutions. *Protein Sci*, 11(6):1393–1408, 2002.

[135] S. C. Lovell, I. W. Davis, W. B. Arendall, P. I. W. de Bakker, J. M. Word, M. G. Prisant, J. S. Richardson, and D. C. Richardson. Structure validation by Calpha geometry: phi,psi and Cbeta deviation. *Proteins*, 50(3):437–450, 2003.

[136] P.l. D. Lyne. Structure-based virtual screening: an overview. *Drug Discov Today*, 7(20):1047–1055, Oct 2002.

[137] A. D. MacKerell, Jr., D. Bashford, M. Bellott, R. L. Dunbrack Jr., J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, III, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B*, 102:3586–3616, 1998.

[138] L. Marinelli, K. E. Gottschalk, A. Meyer, E. Novellino, and H. Kessler. Human integrin alpha v beta 5: Homology modeling and ligand binding. *J Med Chem*, 47(17):4166–4177, 2004.

[139] M. R. Mcgann, H. R. Almond, A. Nicholls, J. A. Grant, and F. K. Brown. Gaussian docking functions. *Biopolymers*, 68(1):76–90, 2003.

[140] C. McMartin and R. S. Bohacek. QXP: powerful, rapid computer algorithms for structure-based drug design. *J Comput Aided Mol Des*, 11(4):333–344, Jul 1997.

[141] J. Mendes, A. M. Baptista, M. A. Carrondo, and C. M. Soares. Improved modeling of side-chains in proteins with rotamer-based methods: a flexible rotamer model. *Proteins*, 37(4):530–543, 1999.

[142] E. C. Meng, B. K. Shoichet, and I. D. Kuntz. Automated docking with grid-based energy evaluation. *J Comput Chem*, 13(4):505–524, 1992.

[143] J. B. O. Mitchell, R. A. Laskowski, A. Alex, and J. M. Thornton. BLEEP - potential of mean force describing protein-ligand interactions: I. generating potential. *J Comp Chem*, 20:1165–1176, 1999.

[144] J. B. O. Mitchell, R. A. Laskowski, A. Alex, and J. M. Thornton. BLEEP - potential of mean force describing protein-ligand interactions: II. calculation of binding energies and comparison with experimental data. *J Comp Chem*, 20:1177–1185, 1999.

[145] M. A. Miteva, W. H. Lee, M. O. Montes, and B. O. Villoutreix. Fast structure-based virtual ligand screening combining FRED, DOCK, and Surflex. *J Med Chem*, 48(19):6012–6022, 2005.

[146] V. Mohan, A. C. Gibbs, M. D. Cummings, E. P. Jaeger, and R. L. DesJarlais. Docking: successes and challenges. *Curr Pharm Des*, 11(3):323–333, 2005.

[147] G. M. Morris, D. S. Goodsel, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, and A. J. Olson. Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *J Comp Chem*, 19:1639–1662, 1998.

[148] J. Moult, K. Fidelis, B. Rost, T. Hubbard, and A. Tramontano. Critical assessment of methods of protein structure prediction (casp) - round 6. *Proteins*, 61 Suppl 7:3–7, 2005.

[149] J. Moult, K. Fidelis, A. Zemla, and T. Hubbard. Critical assessment of methods of protein structure prediction CASP-round V. *Proteins*, 53 Suppl 6:334–339, 2003.

[150] I. Muegge. Effect of ligand volume correction on pmf scoring. *J Comput Chem*, 4:418–425, 2001.

[151] I. Muegge and Y. C. Martin. A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J Med Chem*, 42(5):791–804, 1999.

[152] M. Nowak, M. Kolaczkowski, M. Pawlowski, and A. J. Bojarski. Homology Modeling of the Serotonin 5-HT(1A) Receptor Using Automated Docking of Bioactive Compounds with Defined Geometry. *J Med Chem*, 49(1):205–214, 2006.

[153] C. Oshiro, E. K. Bradley, J. Eksterowicz, E. Evensen, M. L. Lamb, J. K. Lanctot, S. Putta, R. Stanton, and P. D. J. Grootenhuis. Performance of 3D-database molecular docking studies into homology models. *J Med Chem*, 47(3):764–767, 2004.

[154] F. Osterberg, G. M. Morris, M. F. Sanner, A. J. Olson, and D. S. Goodsell. Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock. *Proteins*, 46(1):34–40, 2002.

[155] B. Park and M. Levitt. Energy functions that discriminate X-ray and near native folds from well-constructed decoys. *J Mol Biol*, 258(2):367–392, 1996.

[156] S. Parthasarathy and M. R. Murthy. Analysis of temperature factor distribution in high-resolution protein structures. *Protein Sci*, 6(12):2561–2567, 1997.

[157] D. A. Pearlman, D. A. Case, J. W. Caldwell, W. S. Ross, T. E. Cheatham, S. Debolt, D. Ferguson, G. Seibel, and P. Kollman. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comput Phys Commun*, 91(1-3):1–41, 1995.

[158] M. C. Peitsch, T. Schwede, and N. Guex. Automated protein modelling - the proteome in 3D. *Pharmacogenomics*, 1(3):257–266, 2000.

[159] R. W. Peterson, P. L. Dutton, and A. J. Wand. Improved side-chain prediction accuracy using an ab initio potential energy function and a very large rotamer library. *Protein Sci*, 13(3):735–751, 2004.

[160] N. A. Pierce and E. Winfree. Protein design is NP-hard. *Protein Eng*, 15(10):779–782, 2002.

[161] A. Poleksic, J. F. Danzer, B. A. Palmer, B. D. Olafson, and D. A. Debe. SPINFAST: using structure alignment profiles to enhance the accuracy and assess the reliability of protein side-chain modeling. *Proteins*, 65(4):953–958, 2006.

[162] T. Polgár and G. M. Keserü. Ensemble docking into flexible active sites. critical evaluation of FlexE against JNK-3 and beta-secretase. *J Chem Inf Model*, 46(4):1795–1805, 2006.

[163] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical recipes in C: The art of scientific computing.* Cambridge University Press, 1986.

[164] J. R. Quinlan. *C4.5: programs for machine learning.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[165] J. R. Quinlan. *Induction of Decision Trees*, volume 1. Kluwer Academic Publishers, Hingham, MA, USA, 2003.

[166] J. Rahnenführer. personal communication. 2006.

[167] G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.*, 7:95–9, 1963.

[168] M. Rarey. *Flexx Release 2 User Guide.* BioSolveIT GmbH, An der Ziegelei 75, 53757 Sankt Augustin, Germany, 2007.

[169] M. Rarey, J. Degen, and I. Reulecke. *Docking and Scoring for Structure-based Drug Design. In: Bioinformatics - From Genome to Therapies, Volume 2, Ed.: T. Lengauer.* Wiley-VCH Verlag, 2007.

[170] M. Rarey and J. S. Dixon. Feature trees: A new molecular similarity measure based on tree matching. *J Comput Aided Mol Des*, 12:471–490, 1998.

[171] M. Rarey, B. Kramer, and T. Lengauer. Multiple automatic base selection: protein-ligand docking based on incremental construction without manual intervention. *J Comput Aided Mol Des*, 11(4):369–384, 1997.

[172] M. Rarey, B. Kramer, T. Lengauer, and G. Klebe. A fast flexible docking method using an incremental construction algorithm. *J Mol Biol*, 261(3):470–489, 1996.

[173] M. Rarey, S. Wefing, and T. Lengauer. Placement of medium-sized molecular fragments into active sites of proteins. *J Comput Aided Mol Des*, 10(1):41–54, 1996.

[174] H. W. Reesink, S. Zeuzem, C. J. Weegink, N. Forestier, A. van Vliet, J. van de Wetering de Rooij, L. McNair, S. Purdy, R. Kauffman, J. Alam, and P. L. M. Jansen. Rapid decline of viral RNA in hepatitis C patients treated with VX-950: a phase Ib, placebo-controlled, randomized study. *Gastroenterology*, 131(4):997–1002, 2006.

[175] G. Rhodes. Judging the quality of macromolecular models - a glossary of terms from crystallography, NMR, and homology modeling. Published online on http://www.usm.maine.edu/∼rhodes/ModQual/, 2007.

[176] Gale Rhodes. *Crystallography Made Crystal Clear*. Academic Press, 2000.

[177] F. M. Richards. Areas, volumes, packing and protein structure. *Annu Rev Biophys Bioeng*, 6:151–176, 1977.

[178] J. S. Richardson. The anatomy and taxonomy of protein structure. *Adv Protein Chem*, 34:167–339, 1981.

[179] R. N. Riemann and M. Zacharias. Refinement of protein cores and protein-peptide interfaces using a potential scaling approach. *Protein Eng Des Sel*, 18(10):465–476, 2005.

[180] J. Sadowski, J. Gasteiger, and G. Klebe. Comparison of automatic three-dimensional model builders using 639 x-ray structures. *J Chem Inf Comput Sci*, 34:1000–1008, 1994.

[181] A. Sali and T. L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*, 234(3):779–815, 1993.

[182] A. Sali and J. P. Overington. Derivation of rules for comparative protein modeling from a database of protein structure alignments. *Protein Sci*, 3(9):1582–1596, 1994.

[183] R. Samudrala and M. Levitt. Decoys 'R' Us: a database of incorrect conformations to improve protein structure prediction. *Protein Sci*, 9(7):1399–1401, 2000.

[184] R. Samudrala and J. Moult. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J Mol Biol*, 275(5):895–916, 1998.

[185] R. Samudrala and J. Moult. Determinants of side chain conformational preferences in protein structures. *Protein Eng*, 11(11):991–997, 1998.

[186] C. Sarrazin, T. L. Kieffer, D. Bartels, B. Hanzelka, U. Müh, M. Welker, D. Wincheringer, Y. Zhou, H. M. Chu, C. Lin, C. Weegink, H. Reesink, S. Zeuzem, and A. D. Kwong. Dynamic hepatitis C virus genotypic and phenotypic changes in patients treated with the protease inhibitor telaprevir. *Gastroenterology*, 132(5):1767–1777, 2007.

[187] A. Schafferhans and G. Klebe. Docking ligands onto binding site representations derived from proteins built by homology modelling. *J Mol Biol*, 307(1):407–427, 2001.

[188] T. Schlick. *Molecular Modeling and Simulation - An Interdisciplinary Guide.* Springer, 2002.

[189] V. Schnecke and L. A. Kuhn. Virtual screening with solvation and ligand-induced complementarity. *Persp Drug Disc Design*, 20:171–190, 2000.

[190] H. Schrauber, F. Eisenhaber, and P. Argos. Rotamers: to be or not to be? an analysis of amino acid side-chain conformations in globular proteins. *J Mol Biol*, 230(2):592–612, 1993.

[191] T. Schwede, J. Kopp, N. Guex, and M. C. Peitsch. SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res*, 31(13):3381–3385, 2003.

[192] P. S. Shenkin, H. Farid, and J. S. Fetrow. Prediction and evaluation of side-chain conformations for protein backbone structures. *Proteins*, 26(3):323–352, 1996.

[193] W. Sherman, T. Day, M. P. Jacobson, R. A. Friesner, and R. Farid. Novel procedure for modeling ligand/receptor induced fit effects. *J Med Chem*, 49(2):534–553, 2006.

[194] R. P. Shetty, P. I. W. De Bakker, M. A. DePristo, and T. L. Blundell. Advantages of fine-grained side chain conformer libraries. *Protein Eng*, 16(12):963–969, 2003.

[195] B. K. Shoichet and I. D. Kuntz. Protein docking and complementarity. *J Mol Biol*, 221(1):327–346, 1991.

[196] M. J. Sippl. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol*, 213(4):859–883, 1990.

[197] M. J. Sippl. Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J Comput Aided Mol Des*, 7(4):473–501, 1993.

[198] M. J. Slater, E. M. Amphlett, D. M. Andrews, P. Bamborough, S. J. Carey, M. R. Johnson, P. S. Jones, G. Mills, N. R. Parry, D. O'N. Somers, A. J. Stewart, and T. Skarzynski. Pyrrolidine-5,5-trans-lactams. 4. incorporation of a P3/P4 urea leads to potent intracellular inhibitors of hepatitis C virus NS3/4A protease. *Org Lett*, 5(24):4627–4630, 2003.

[199] C. A. Sotriffer, H. Gohlke, and G. Klebe. Docking into knowledge-based potential fields: a comparative evaluation of drugscore. *J Med Chem*, 45(10):1967–1970, 2002.

[200] M. Stahl and M. Rarey. Detailed analysis of scoring functions for virtual screening. *J Med Chem*, 44(7):1035–1042, 2001.

[201] A. Steffen, J. Günther, and H. Briem. Evaluation of the applicability of the flexe ensemble docking approach to virtual screening for CDK2 inhibitors. 18. Darmstädter Molecular Modelling Workshop., 2004.

[202] C. M. Summa, M. Levitt, and W. F. Degrado. An atomic environment potential for use in protein structure prediction. *J Mol Biol*, 352(4):986–1001, 2005.

[203] Tripos Inc. SYBYL 7.0. 1699 South Hanley Rd., St. Louis, Missouri, 63144.

[204] M. L. Teodoro and L. E. Kavraki. Conformational flexibility models for the receptor in structure based drug design. *Curr Pharm Des*, 9(20):1635–1648, 2003.

[205] J. B. Thoden, T. M. Wohlers, J. L. Fridovich-Keil, and H. M. Holden. Crystallographic evidence for Tyr 157 functioning as the active site base in human UDP-galactose 4-epimerase. *Biochemistry*, 39(19):5691–5701, 2000.

[206] P. D. Thomas and K. A. Dill. An iterative method for extracting energy-like quantities from protein structures. *Proc Natl Acad Sci U S A*, 93(21):11628–11633, 1996.

[207] P. D. Thomas and K. A. Dill. Statistical potentials extracted from protein structures: how accurate are they? *J Mol Biol*, 257(2):457–469, 1996.

[208] H. B. Thorsteinsdottir, T. Schwede, V. Zoete, and M. Meuwly. How inaccuracies in protein structure models affect estimates of protein-ligand interactions: computational analysis of HIV-I protease inhibitor binding. *Proteins*, 65(2):407–423, 2006.

[209] S. C. E. Tosatto. The victor/frst function for model quality estimation. *J Comput Biol*, 12(10):1316–1327, 2005.

[210] P. Tufféry, C. Etchebest, and S. Hazout. Prediction of protein side chain conformations: a study on the influence of backbone accuracy on conformation stability in the rotamer space. *Protein Eng*, 10(4):361–372, 1997.

[211] P. Tufféry, C. Etchebest, S. Hazout, and R. Lavery. A new approach to the rapid determination of protein side chain conformations. *J Biomol Struct Dyn*, 8(6):1267–1289, 1991.

[212] S. Ulmschneider, U. Müller-Vieira, C. D. Klein, I. Antes, T. Lengauer, and R. W. Hartmann. Synthesis and evaluation of (pyridylmethylene)tetrahydronaphthalenes/-indanes and structurally modified derivatives: potent and selective inhibitors of aldosterone synthase. *J Med Chem*, 48(5):1563–1575, 2005.

[213] S. Ulmschneider, U. Müller-Vieira, M. Mitrenga, R. W. Hartmann, S. Oberwinkler-Marchais, C. D. Klein, M. Bureik, R. Bernhardt, I. Antes, and T. Lengauer. Synthesis and evaluation of imidazolylmethylenetetrahydronaphthalenes and imidazolylmethyleneindanes: potent inhibitors of aldosterone synthase. *J Med Chem*, 48(6):1796–1805, 2005.

[214] H. F. G. Velec, H. Gohlke, and G. Klebe. DrugScore(CSD)-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *J Med Chem*, 48(20):6296–6303, 2005.

[215] C. M. Venkatachalam, X. Jiang, T. Oldfield, and M. Waldman. LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *J Mol Graph Model*, 21(4):289–307, Jan 2003.

[216] B. Wallner and A. Elofsson. All are not equal: a benchmark of different homology modeling programs. *Protein Sci*, 14(5):1315–1327, 2005.

[217] B. Wallner, P. Larsson, and A. Elofsson. Pcons.net: protein structure prediction meta server. *Nucleic Acids Res*, 35(Web Server issue):W369–W374, 2007.

[218] C. Wang, O. Schueler-Furman, and D. Baker. Improved side-chain modeling for protein-protein docking. *Protein Sci*, 14(5):1328–1339, 2005.

[219] G. Wang and R. L. Dunbrack. PISCES: a protein sequence culling server. *Bioinformatics*, 19(12):1589–1591, 2003.

[220] G. Wang and R. L. Dunbrack. PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Res*, 33(Web Server issue):W94–W98, 2005.

[221] J. M. Wang, P. Cieplak, and P. A. Kollman. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J Comput Chem*, 21:1049–1074, 2000.

[222] R. Wang, L. Lai, and S. Wang. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J Comput Aided Mol Des*, Volume 16(1):11–26, 2002.

[223] R. Wang, Y. Lu, and S. Wang. Comparative evaluation of 11 scoring functions for molecular docking. *J Med Chem*, 46(12):2287–2303, 2003.

[224] G. L. Warren, C. W. Andrews, A. M. Capelli, B. Clarke, J. LaLonde, M. H. Lambert, M. Lindvall, N. Nevins, S. F. Semus, S. Senger, G. Tedesco, I. D. Wall, J. M. Woolven, C. E. Peishoff, and M. S. Head. A critical assessment of docking programs and scoring functions. *J Med Chem*, 49(20):5912–5931, 2006.

[225] B. Q. Wei, L. H. Weaver, A. M. Ferrari, B. W. Matthews, and B. K. Shoichet. Testing a flexible-receptor docking algorithm in a model binding site. *J Mol Biol*, 337(5):1161–1182, 2004.

[226] C. X. Weichenberger and M. J. Sippl. Self-consistent assignment of asparagine and glutamine amide rotamers in protein crystal structures. *Structure*, 14(6):967–972, 2006.

[227] S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Jr. Profeta, and P. K. Weiner. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J Am Chem Soc*, 106:765–784, 1984.

[228] S. J. Weiner, P. A. Kollman, D. T. Nguyen, and D. A. Case. An all atom force field for simulations of proteins and nucleic acids. *J Comp Chem*, 7:230–252, 1986.

[229] C. Welsch, F. Domingues, S. Susser, I. Antes, C. Hartmann, G. Mayr, A. Schlicker, C. Sarrazin, M. Albrecht, S. Zeuzem, and T. Lengauer. Molecular basis of telaprevir resistance due to V36 and T54 mutations in the NS3-4 A protease of HCV. *Genome Biol*, 9(1):R16, Jan 2008.

[230] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2nd edition, 2005.

[231] J. M. Word, S. C. Lovell, J. S. Richardson, and D. C. Richardson. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J Mol Biol*, 285(4):1735–1747, 1999.

[232] S. Wu, J. Skolnick, and Y. Zhang. Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biol*, 5:17, 2007.

[233] Z. Xiang and B. Honig. Extending the accuracy limits of prediction for side-chain conformations. *J Mol Biol*, 311(2):421–430, 2001.

[234] W. Xie and N. V. Sahinidis. Residue-rotamer-reduction algorithm for the protein side-chain conformation problem. *Bioinformatics*, 22(2):188–194, 2006.

[235] Z. Yuan, T. L. Bailey, and R. D. Teasdale. Prediction of protein B-factor profiles. *Proteins*, 58(4):905–912, 2005.

[236] Z. Yuan, J. Zhao, and Z. X. Wang. Flexibility analysis of enzyme active sites by crystallographic temperature factors. *Protein Eng*, 16(2):109–114, 2003.

[237] M. I. Zavodszky and L. A. Kuhn. Side-chain flexibility in protein-ligand binding: the minimal rotation hypothesis. *Protein Sci*, 14(4):1104–1114, 2005.

[238] C. Zhang, S. Liu, H. Y. Zhou, and Y. Q. Zhou. The dependence of all-atom statistical potentials on structural training database. *Biophys J*, 86(6):3349–3358, 2004.

[239] C. Zhang, S. Liu, and Y. Zhou. Accurate and efficient loop selections by the DFIRE-based all-atom statistical potential. *Protein Sci*, 13(2):391–399, 2004.

[240] C. Zhang, S. Liu, Q. Zhu, and Y. Zhou. A knowledge-based energy function for protein-ligand, protein-protein, and protein-dna complexes. *J Med Chem*, 48(7):2325–2335, 2005.

[241] Y. Zhang. Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins*, 69 Suppl 8:108–117, 2007.

[242] Y. Zhang and J. Skolnick. Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc Natl Acad Sci USA*, 101(20):7594–7599, 2004.

[243] Y. Zhao and M. F. Sanner. FLIPDock: Docking flexible ligands into flexible receptors. *Proteins*, 68(3):726–737, 2007.

Appendix A

Docking and Screening Results of the DUD Targets

A.1 Enrichment Plots

Axes

- x - percentage of ligands that were identified by selecting a subset of the ranked compound list
- y - percentage of compounds that were selected from the ranked compound list

Color Key

- black - Enrichment curves for random (lower curve) and optimal (upper curve) selection strategies
- red - FlexX using F-Score and X-ray structures
- cyan - FlexX using ROTA and X-ray structures
- grey - FlexX using F-Score and IRECS models with $\rho_{rot} = 1$
- violet - FlexX using ROTA and IRECS models with $\rho_{rot} = 1$
- blue - FlexE using F-Score and IRECS models with $\rho_{rot} = 2$
- green - FlexE using ROTA and IRECS models with $\rho_{rot} = 3$

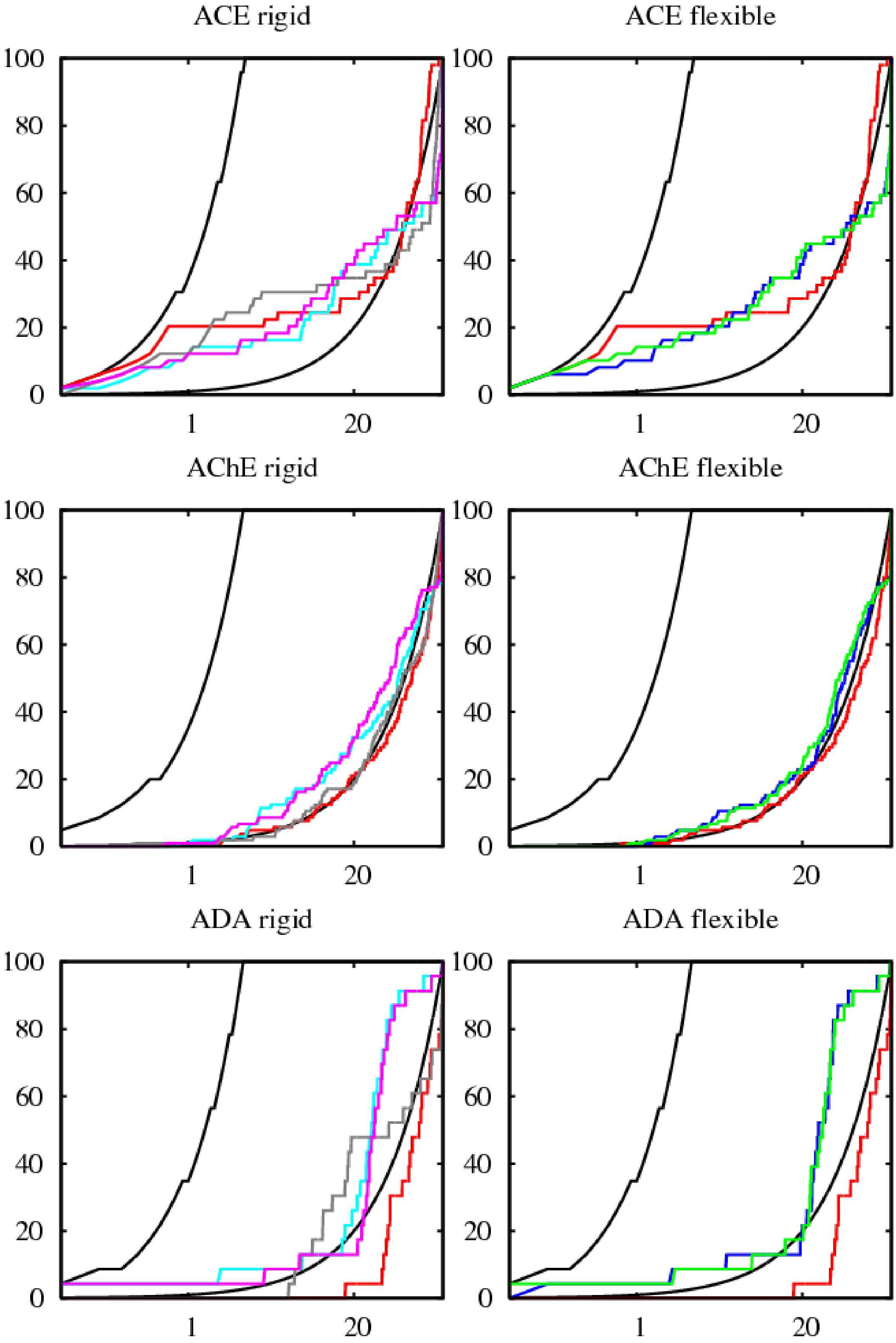


Figure A.1: Enrichment plots for ACE, AChE and ADA

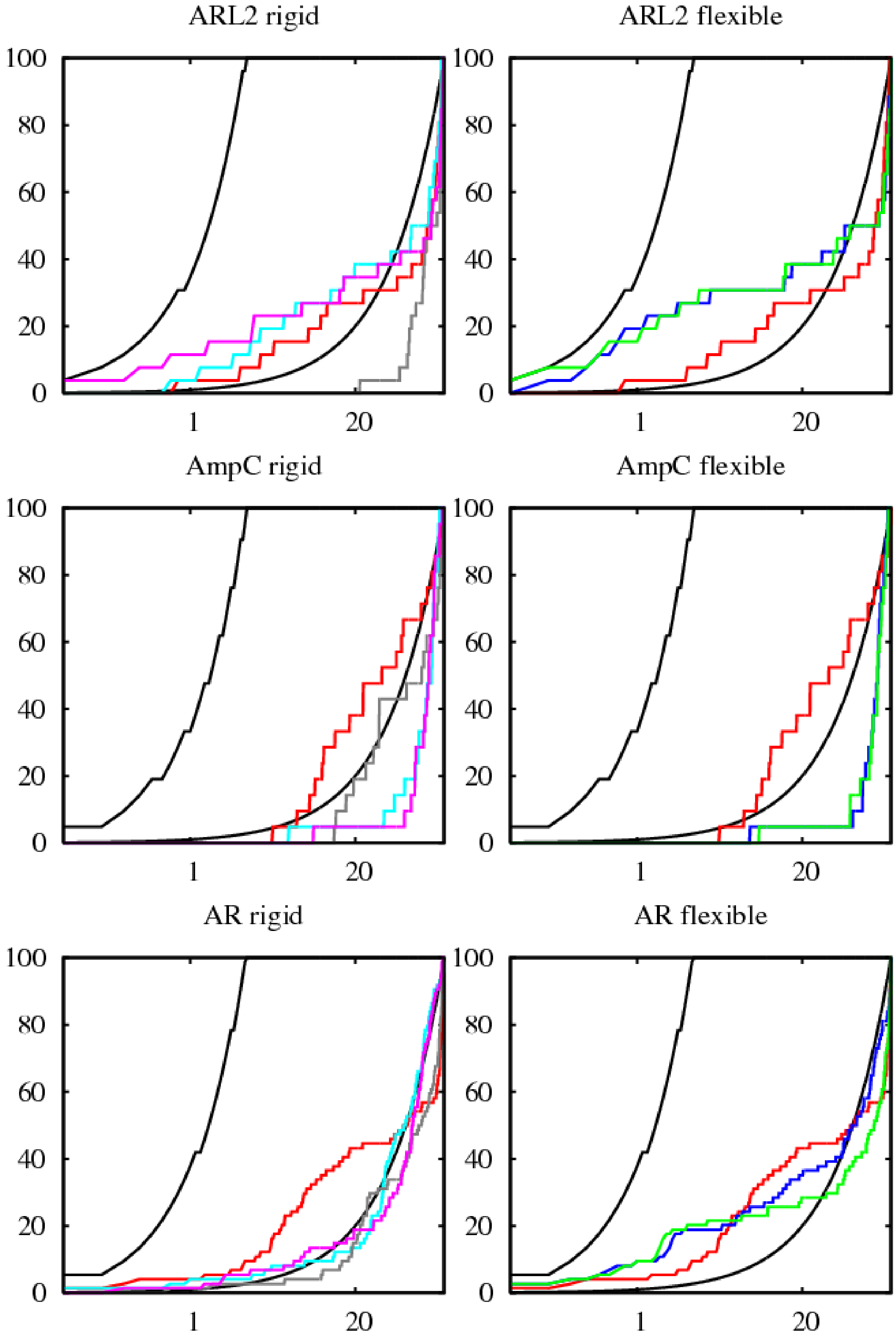


Figure A.2: Enrichment plots for ARL2, AmpC and AR

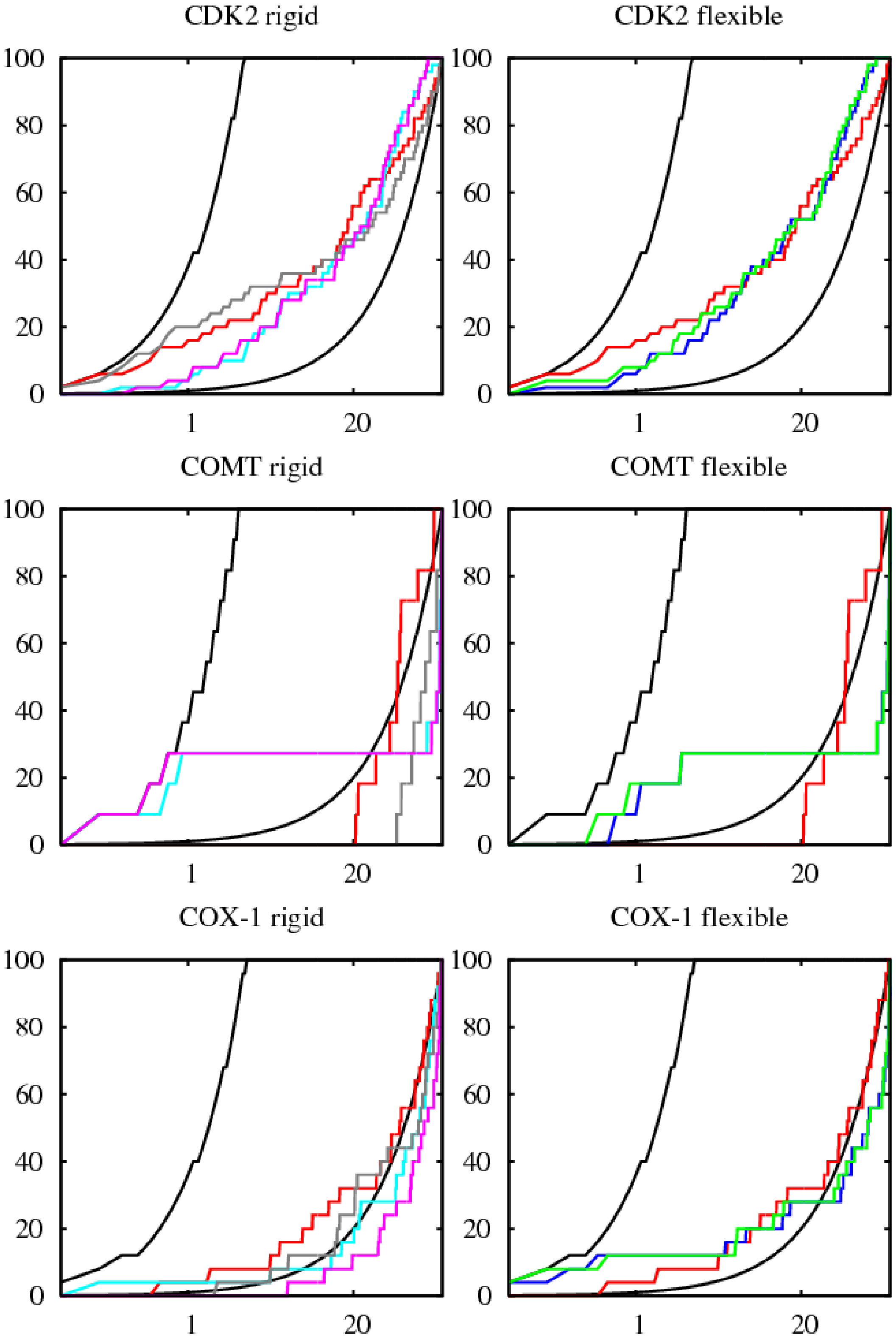


Figure A.3: Enrichment plots for CDK2, COMT and COX-1

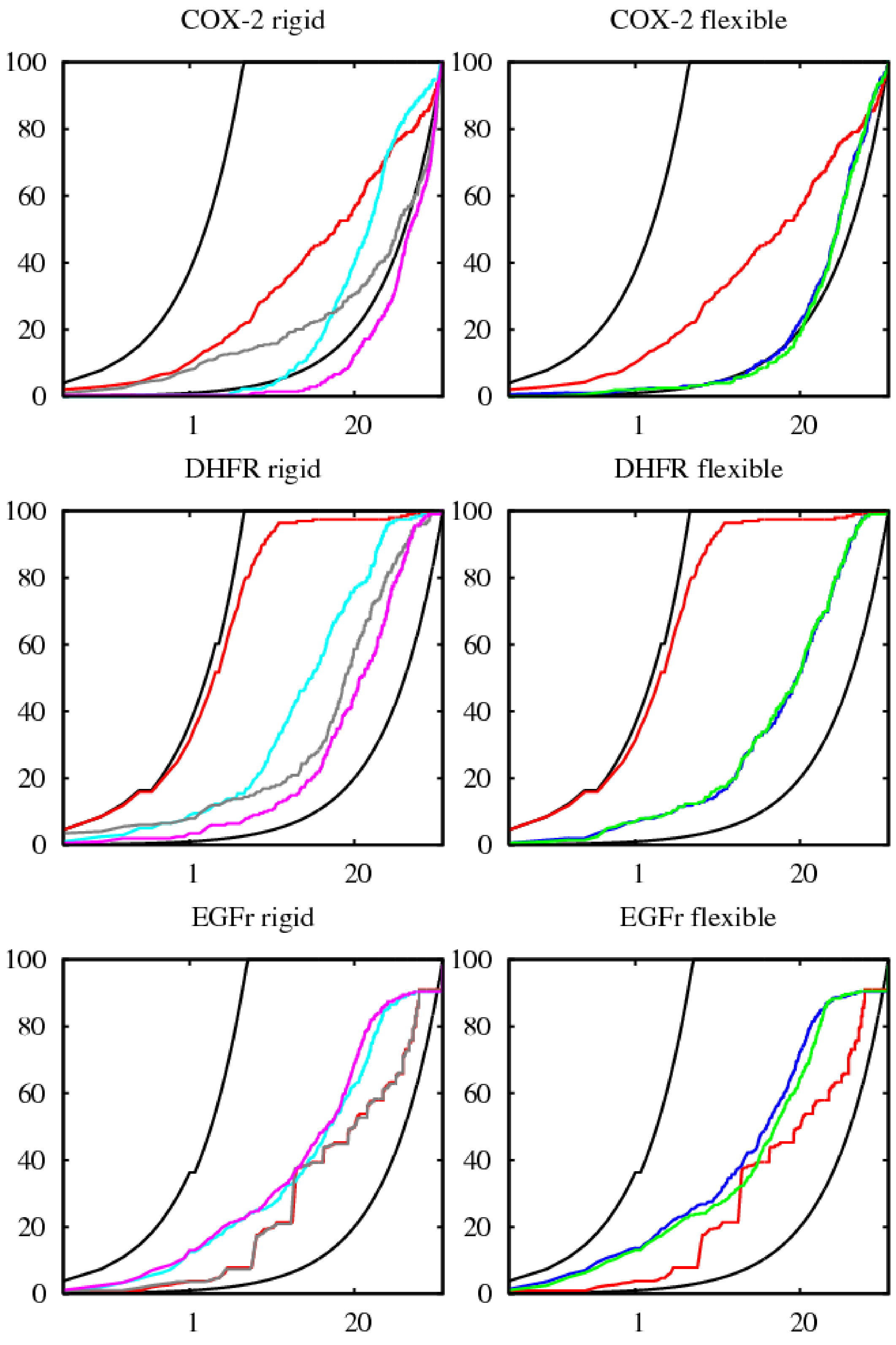


Figure A.4: Enrichment plots for COX-2, DHFR and EGFr

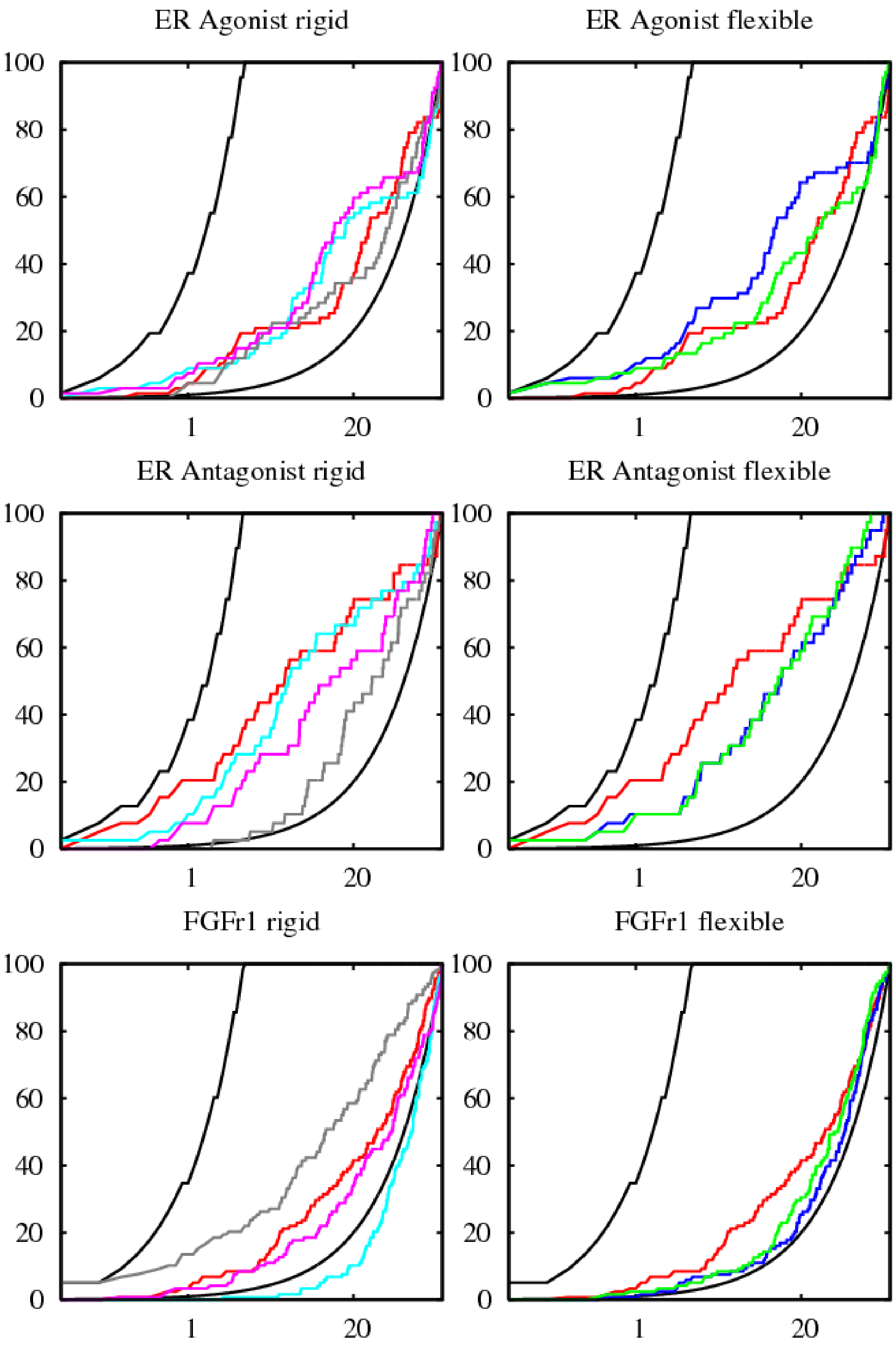


Figure A.5: Enrichment plots for $\mathrm{ER}_{\mathrm{agonist}}$, $\mathrm{ER}_{\mathrm{antagonist}}$ and FGFr1

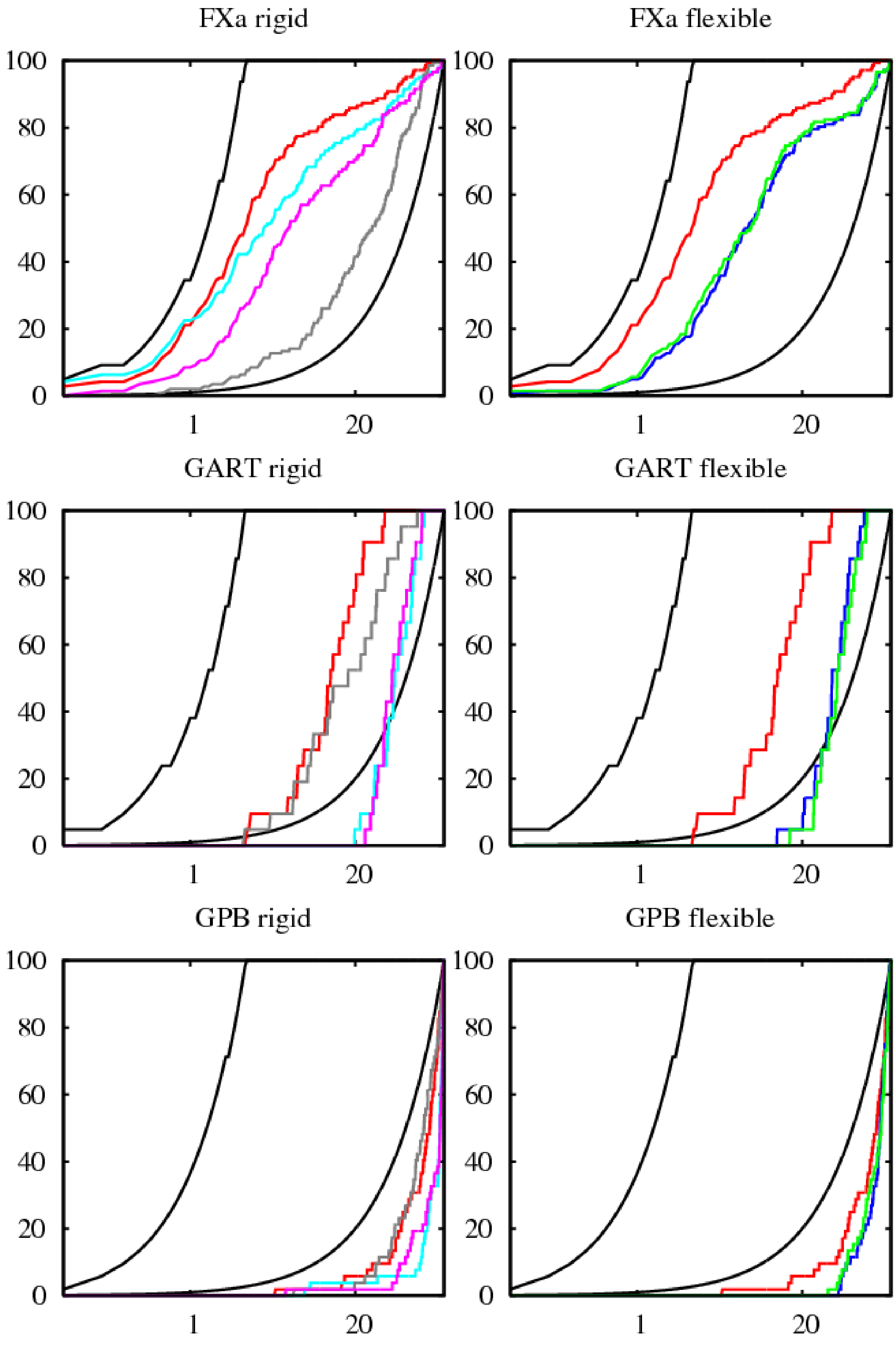


Figure A.6: Enrichment plots for FXa, GART and GPB

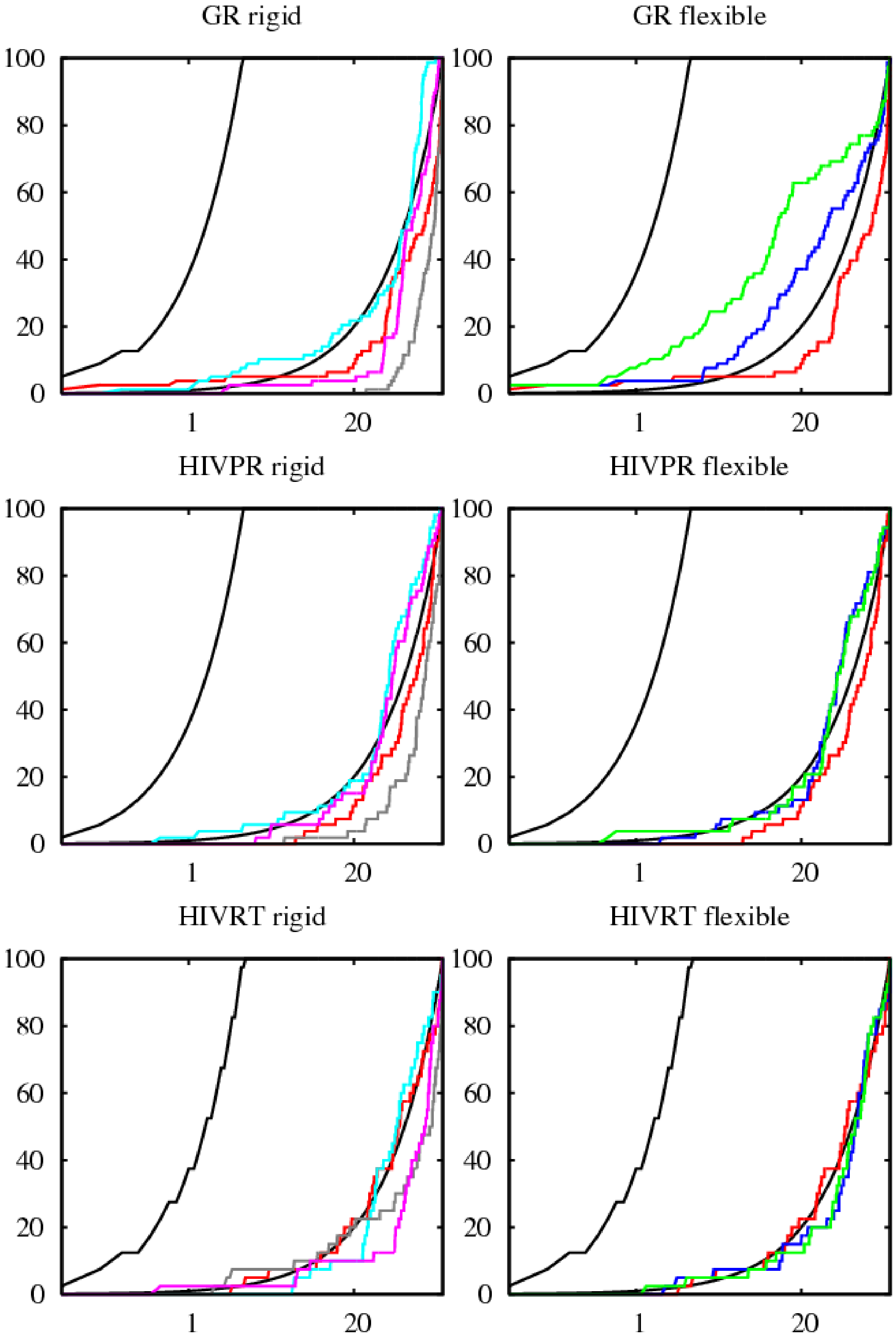


Figure A.7: Enrichment plots for GR, HIVPR and HIVRT

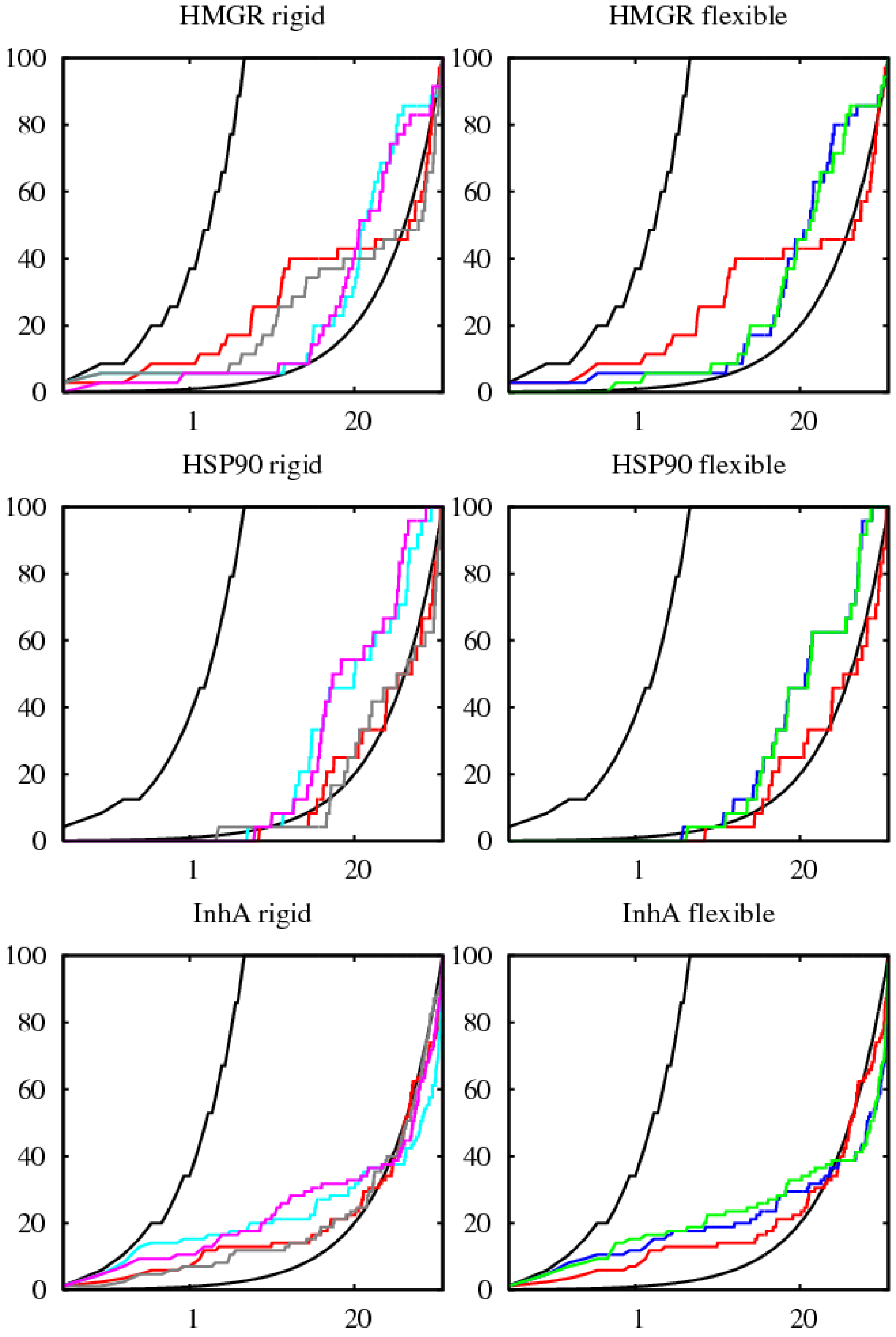


Figure A.8: Enrichment plots for HMGR, HSP90 and InhA

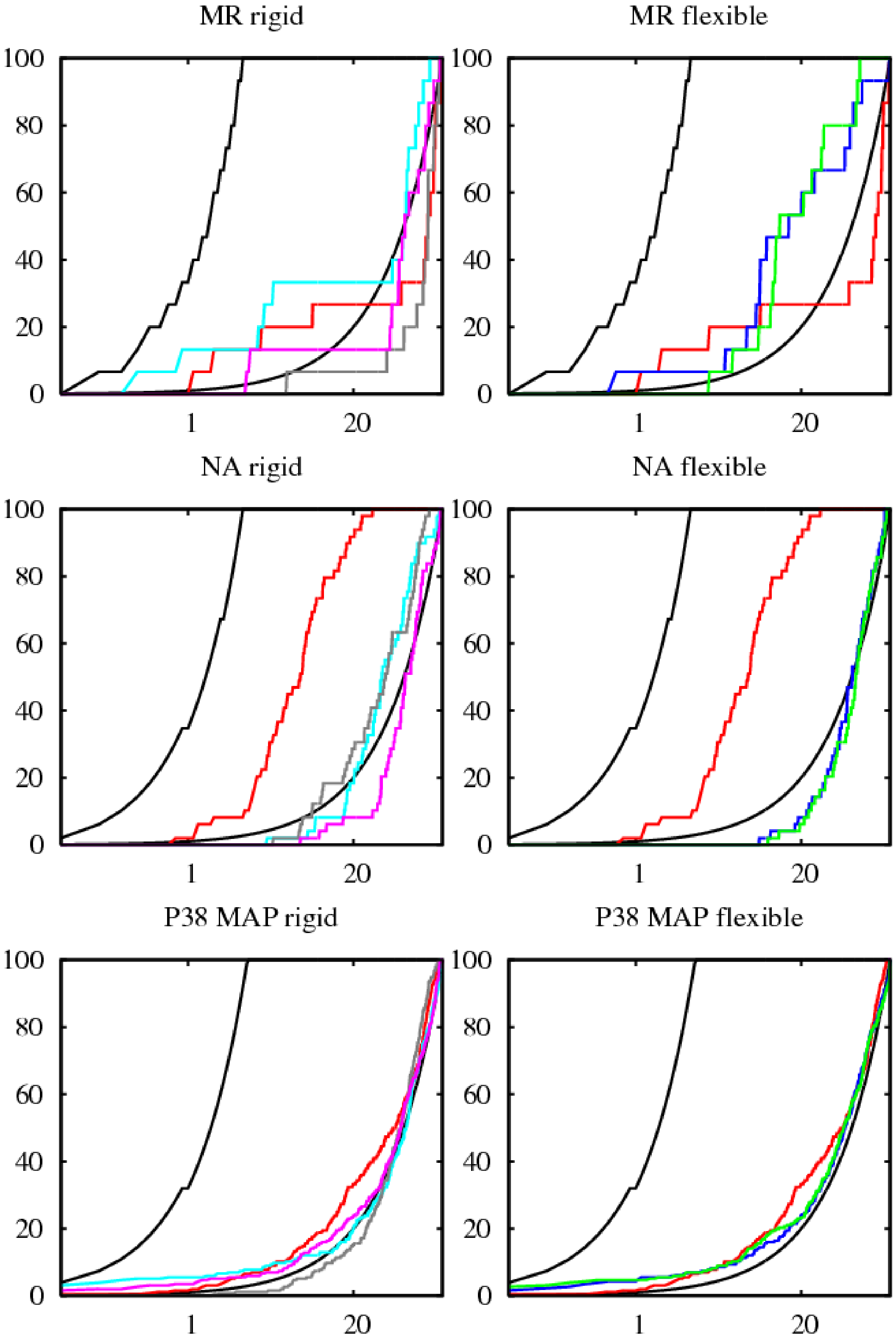
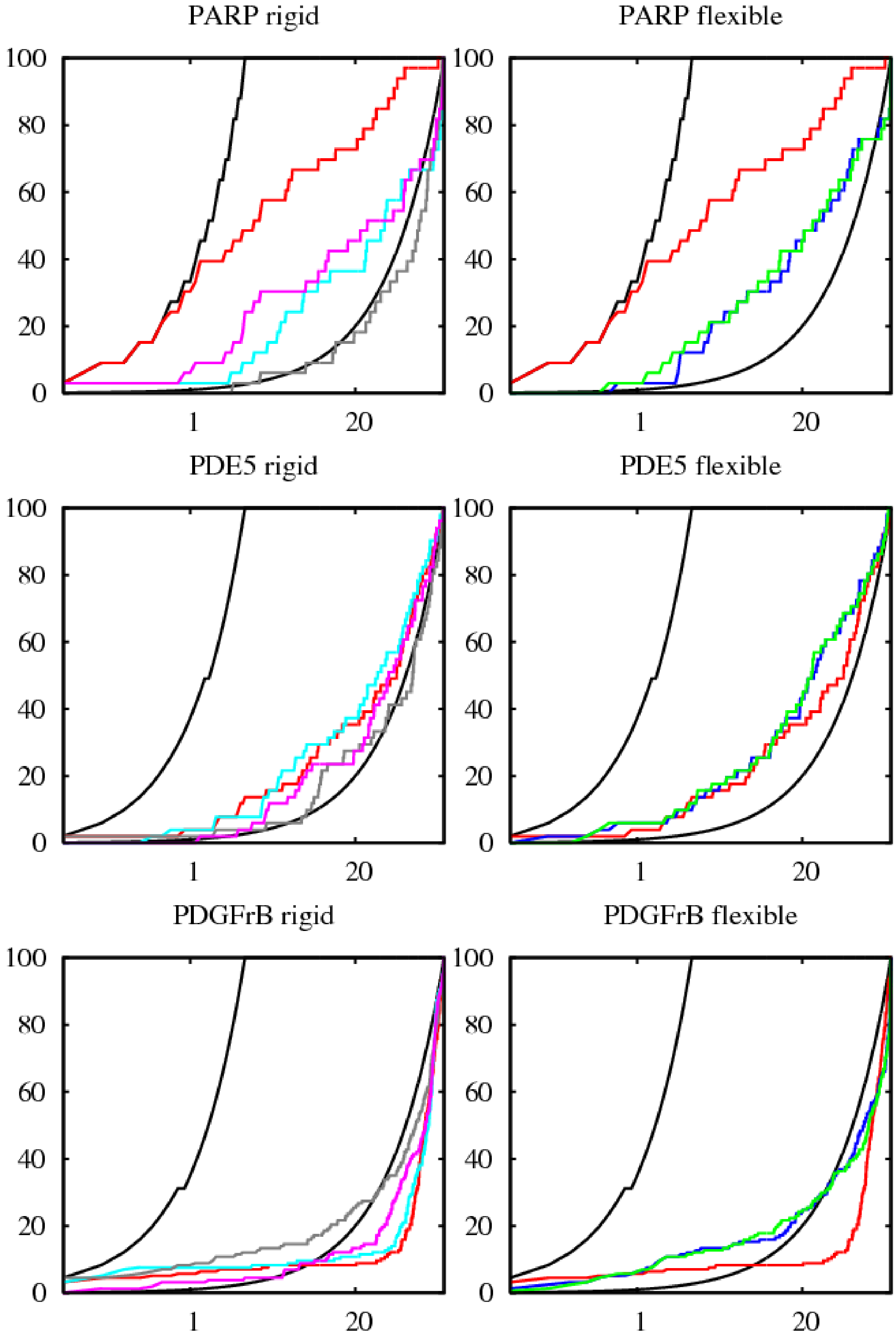


Figure A.9: Enrichment plots for MR, NA and P38 MAP



PARP rigid

PARP flexible

PDE5 rigid

PDE5 flexible

PDGFrB rigid

PDGFrB flexible

Figure A.10: Enrichment plots for PARP, PDE5 and PDGFrB

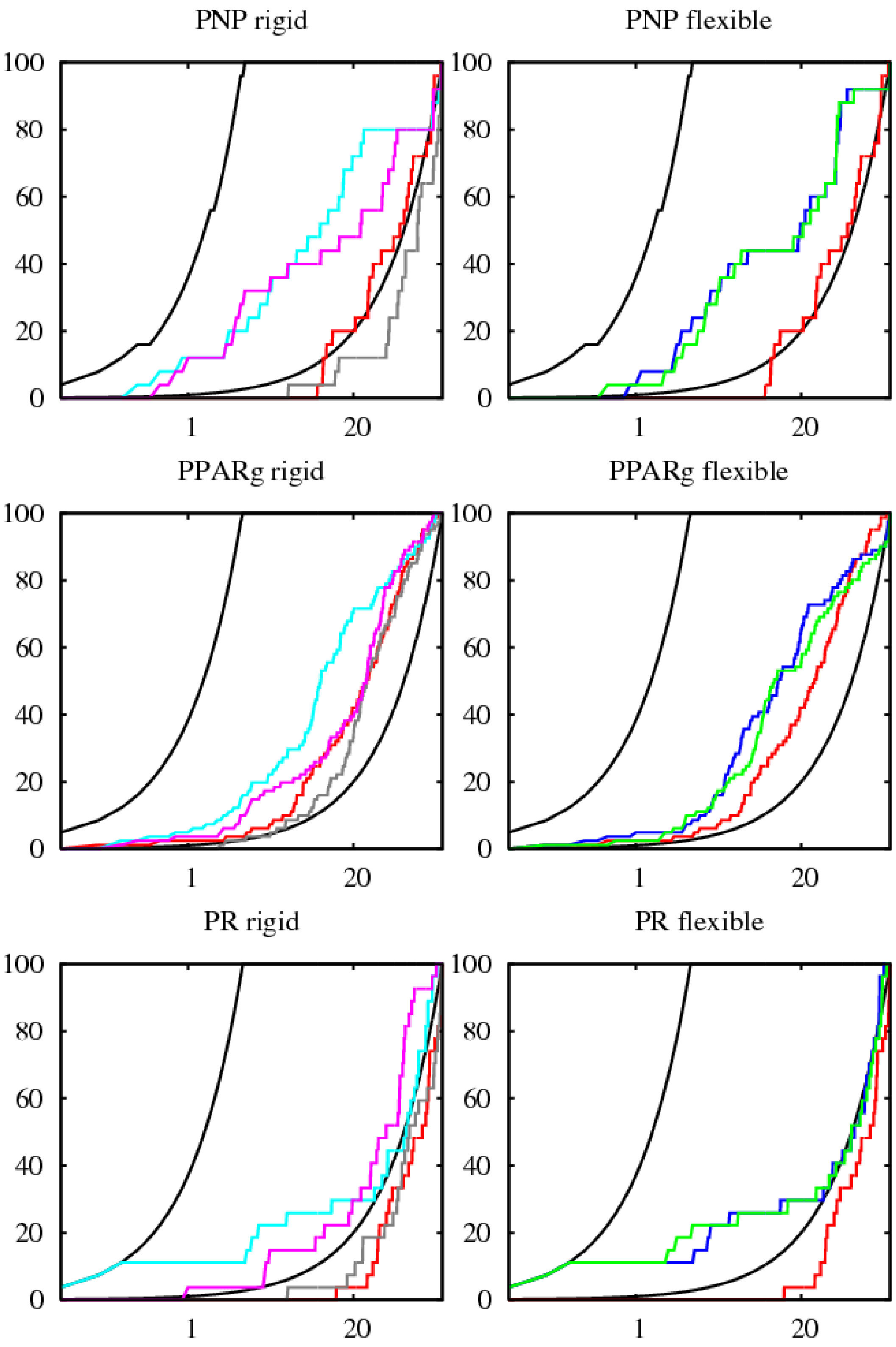


Figure A.11: Enrichment plots for PNP, PPARg and PR

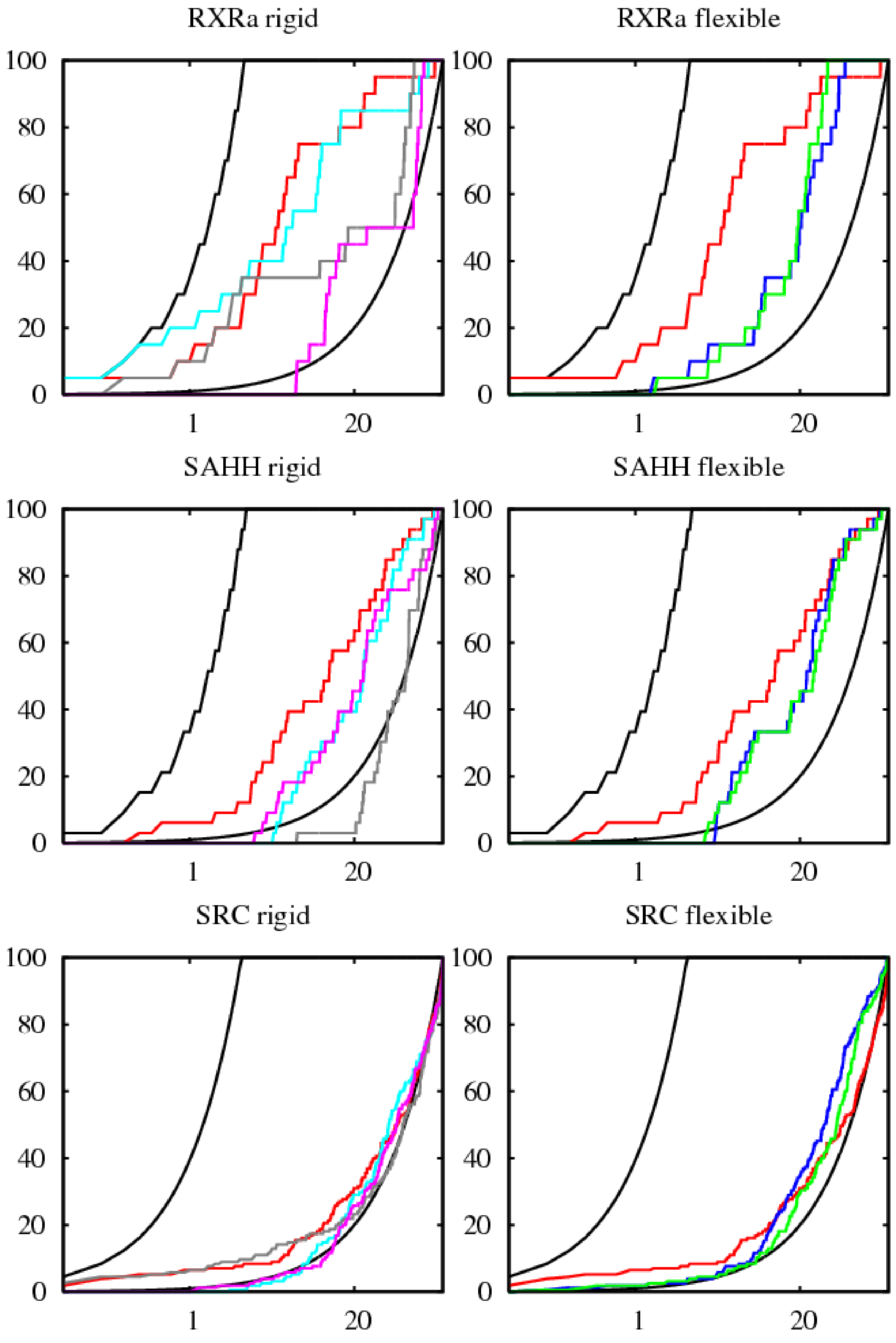


Figure A.12: Enrichment plots for RXRa, SAHH and SRC

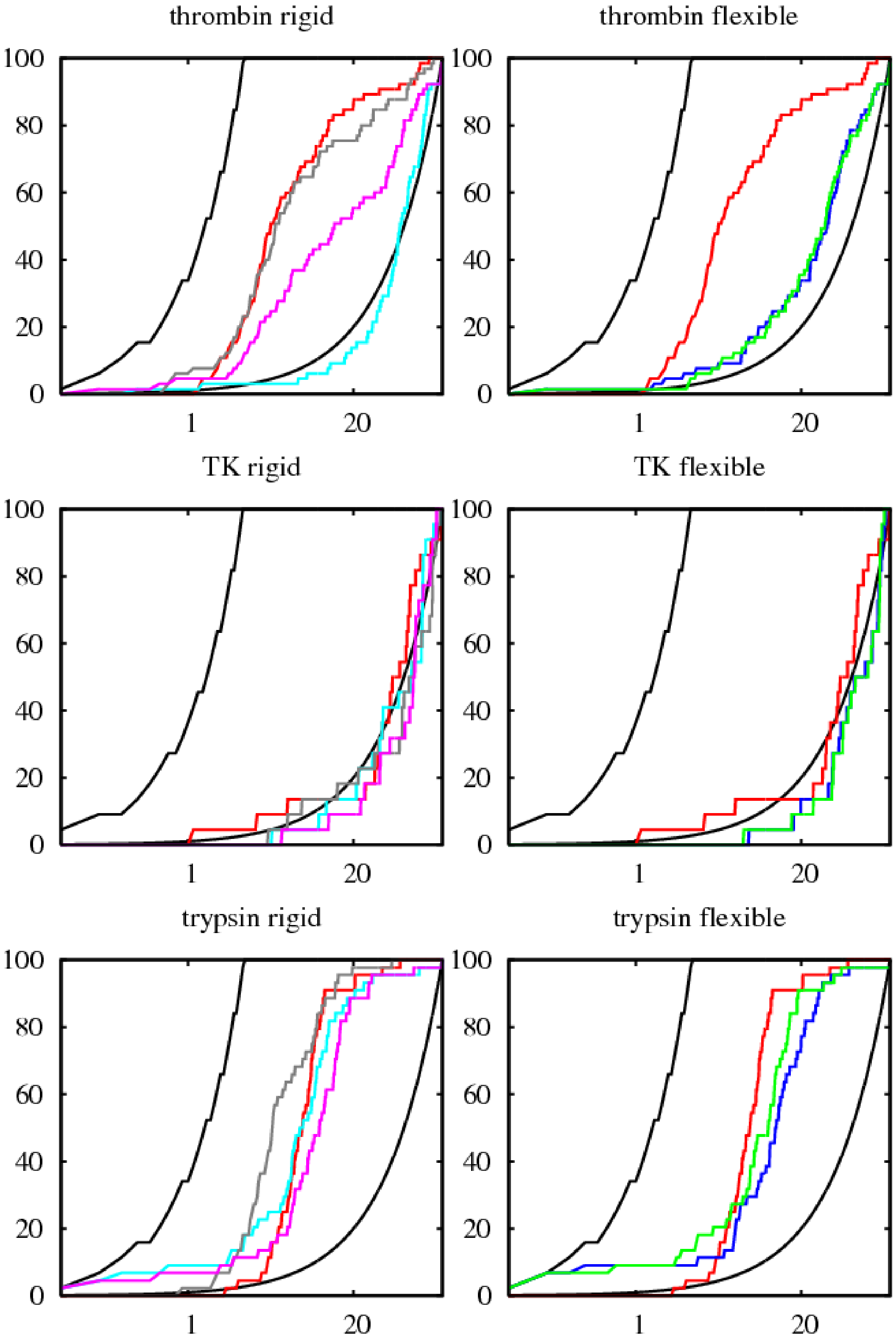


Figure A.13: Enrichment plots for thrombin, TK and trypsin

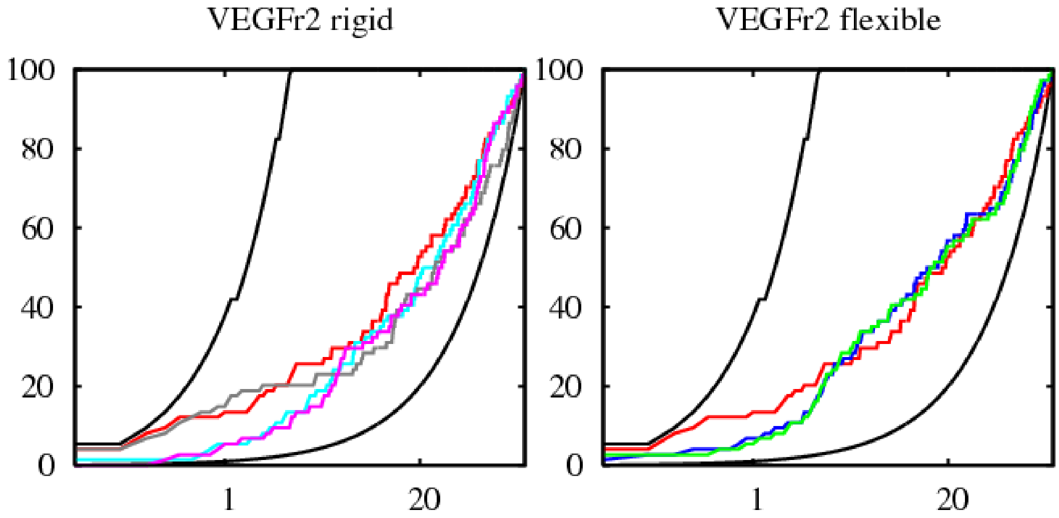


Figure A.14: Enrichment plots for VEGFr2

Appendix B

Mapping between PDB Atom Names and ROTA Atom Types

(a) Alanine

PDB name	ROTA type
N	N.am
O	O
C	C.O
CA	C.H1
CB	C.HX

(b) Arginine

PDB name	ROTA type
N	N.am
O	O
C	C.O
CA	C.H1
CB	C.H2
CG	C.H2
CD	C.H2
NE	N.H1
CZ	C.H0
NH1	N.HX
NH2	N.HX

(c) Asparagine

PDB name	ROTA type
N	N.am
O	O
C	C.O
CA	C.H1
CB	C.H2
CG	C.H0
OD1	O
ND2	N.HX

(d) Aspartatic Acid

PDB name	ROTA type
N	N.am
O	O
C	C.O
CA	C.H1
CB	C.H2
CG	C.O2
OD1	O.CO2
OD2	O.CO2

(e) Cysteine

PDB name	ROTA type
N	N.am
O	O
C	C.O
CA	C.H1
CB	C.H2
SG	S

(f) Glutamine

PDB name	ROTA type
N	N.am
O	O
C	C.O
CA	C.H1
CB	C.H2
CG	C.H2
CD	C.O
OE1	O
NE2	N.HX

(g) Glutamic Acid

PDB name	ROTA type
N	N.am
O	O
C	C.O
CA	C.H1
CB	C.H2
CG	C.H2
CD	C.O2
OE1	O.CO2
OE2	O.CO2

(h) Glycine

PDB name	ROTA type
N	N.am
O	O
C	C.O
CA	C.H1

(i) Histidine

PDB name	ROTA type
N	N.am
O	O
C	C.O
CA	C.H1
CB	C.H2
CG	C.ar
ND1	N.ar
CD2	C.ar
CE1	C.ar
NE2	C.ar

(j) Isoleucine

PDB name	ROTA type
N	N.am
O	O
C	C.O
CA	C.H1
CB	C.H2
CG1	C.H2
CG2	C.HX
CD1	C.HX

(k) Leucine

PDB name	ROTA type
N	N.am
O	O
C	C.O
CA	C.H1
CB	C.H2
CG	C.H1
CD1	C.HX
CD2	C.HX

(l) Lysine

PDB name	ROTA type
N	N.am
O	O
C	C.O
CA	C.H1
CB	C.H2
CG	C.H2
CD	C.H2
CE	C.H2
NZ	N.H3

(m) Methionine

PDB name	ROTA type
N	N.am
O	O
C	C.O
CA	C.H1
CB	C.H2
CG	C.H2
SD	S

(n) Phenylalanine

PDB name	ROTA type
N	N.am
O	O
C	C.O
CA	C.H1
CB	C.H2
CG	C.ar
CD1	C.ar
CD2	C.ar
CE1	C.ar
CE2	C.ar
CZ	C.ar

(o) Proline

PDB name	ROTA type
N	N.am
O	O
C	C.O
CA	C.H1
CB	C.H2
CG	C.H2
CD	C.H2

(p) Serine

PDB name	ROTA type
N	N.am
O	O
C	C.O
CA	C.H1
CB	C.H2
OG	O.H1

(q) Threonine

PDB name	ROTA type
N	N.am
O	O
C	C.O
CA	C.H1
CB	C.H2
OG1	O.H1
CG2	C.HX

(r) Tryptophan

PDB name	ROTA type
N	N.am
O	O
C	C.O
CA	C.H1
CB	C.H2
CG	C.ar
CD1	C.ar
CD2	C.ar
NE1	N.ar
CE2	C.ar
CE3	C.ar
CZ2	C.ar
CZ3	C.ar
CH2	C.ar

(s) Tyrosine

PDB name	ROTA type
N	N.am
O	O
C	C.O
CA	C.H1
CB	C.H2
CG	C.ar
CD1	C.ar
CD2	C.ar
CE1	C.ar
CE2	C.ar
CZ	C.ar

(t) Valine

PDB name	ROTA type
N	N.am
O	O
C	C.O
CA	C.H1
CB	C.H1
CG1	C.HX
CG2	C.HX

Appendix C

Protein Structures for Training and Testing

C.1 IRECS training set

1gv9, 1kpf, 1lmi, 1sh8, 1m55, 1mxr, 1ntv, 1ijy, 1vfy, 1r0m, 1s2o, 1jr8, 1lc5, 1ve1, 1inl, 1brt, 1g2y, 1io0, 1j2r, 1qu9, 1m4i, 1o3y, 1sx5, 1hnj, 1ie9, 1fk5, 1pa7, 1nlq, 1nf9, 1my7, 1cru, 1elk, 1i4u, 1m2d

C.2 IRECS test set A: Single side-chain conformations

1f9v, 1bte, 1jcd, 1kko, 1pin, 1jf4, 1q92, 1oaq, 1k4i, 1jy2, 1w6s, 1m7g, 1wdd, 2b82, 1v4p, 1vef, 1ifr, 1bgf, 1of8, 1ltz, 1uz3, 1gpi, 1f7l, 1f46, 1qwy, 1omr, 1r7j, 1n0q, 1oi0, 1cc8, 1ui0, 1p4o, 1vzi, 1qwo, 1gbs, 1ve4, 1jl1, 1dp7, 1nzj, 1c1k, 1qft, 1idp, 1ds1, 1ql0, 1g6u, 1r6x, 1tp6, 2ew0, 1uq5, 1kr4, 1szh, 1jf8, 1ucd, 2cxv, 1oh0, 1pkh, 1ks8, 1v2x, 1ug6, 2tnf, 1uoy, 1k7j, 1tua, 1sg0, 1n7s, 1uas, 1fd3, 1l7a, 1p1m, 1ezg, 1wfb, 256b, 1rhs, 1gmu, 1sqe, 1h2w, 1kt6, 1t61, 1rtt, 1sqs, 1oyg, 1ijq, 1dfm, 1opd, 1kuf, 1c7k, 1v6s, 1urs, 1es5, 1kyf, 1dj0, 1m22, 1svf, 1flm, 1u4g, 1rya, 1g61, 1wcw, 1roc, 1v30, 1mtp, 1ej0, 1qh5, 1hyo, 1arb, 1w5r, 1ekq, 1qw2, 1pz7, 1utg, 1jg1, 1jyk, 1gd0, 1i52, 8abp, 1jnd, 1mla, 1ew4, 1e5k, 1g2q, 1j0p, 1isu, 1o8b, 1v9y, 1pvm, 1jek, 1ukf, 4eug, 1pp0, 1aap, 1rcq, 1l7l, 1i2t, 1whi, 1fye, 2asb, 1i71, 1jx6, 1kll, 1qs1, 1fp2, 3seb, 1qw9, 1gp0, 1ng6, 1o9g, 1gkp, 1dqz, 1lc0, 1fj2, 1t5b, 1o8x, 1r26, 1ra0, 2a35, 1lr7, 1obo, 1m1f, 1df4, 1cs1

C.3 IRECS test set B: Multiple side-chain conformations

2bw4, 2a13, 1c4o, 1e85, 1w66, 1edm, 1tu9, 1itx, 1h4x, 1ah7, 1g66, 1i1w, 1i0d, 1b8o, 1qxy, 1nyc, 1iua, 1vly, 2f01, 1hd2, 1t7r, 1i24, 1eaq, 2pth, 1nuy, 1s0p, 1mkk, 1lzl, 1t2d, 1jfb, 1k3y, 2bu3, 2bog, 2igd, 2bwq, 1aba, 1tqg, 1tzp, 1u9c, 1o98, 1n8k, 1hx0, 1nqj, 2brf, 1dcs, 1ju2, 1obd, 1d5t, 1nnf, 1qqf, 1lq9, 1w5q, 1atg, 1m1q, 1p4c, 1oaa, 1ijv, 1q6z, 2cov, 1qnr, 1gmx, 2cws, 1o9r, 1lni, 1jni, 7a3h, 1m4l, 1o7j, 1qg8, 1kmv, 1a4i, 1q6o, 1si6, 1isp, 2c5a, 1rk6, 1vh5, 1nof, 2pvb, 1kjl, 1c1d, 1p3c, 1gqi, 2f8a, 2f62, 1iqz, 1tt8, 1jet, 2bk9, 1iom, 1vk1, 1v0w, 1l3k, 1s3c, 1r2q, 1upq, 1rqw, 1psr, 1cip, 1w23, 1kqp, 1lo7, 1unq, 2f22, 1k0m, 1r5y, 1wck, 1irq, 1fx2, 1gyx, 1m9z, 1o2d, 1d4o, 2bzv, 1gs5, 1k5c, 1g2b, 1jl0, 1vyi, 1pbj, 1qtw, 1n55, 3sdh, 1n62, 1o6v, 1ok0, 2akf, 1qlw, 1pko, 1nyt, 1cxq, 1gk9, 1rdq, 1dg6, 2czs, 2mhr, 1mg4, 1q7l, 1ikp, 1h97, 2c4b, 1k3i, 1ssx, 1kv7, 1h1n, 2erb, 1u07, 1vzm, 1knm, 1et1, 1rgz, 1thf, 1pjx, 1mj4, 1fsg, 1q5y, 1gvf, 1mj5, 1rut, 1a6m, 1ufy, 1uwc, 1v5v, 1us0, 1vyr, 1o08, 1mfm, 2a26, 1od6, 1nwa,1bkr, 1gci, 1gpp, 1sjy, 1e7l, 1vr7, 1k55, 1u2h, 2eng, 1k3x, 2bry, 1tuk, 2axw, 1ls1, 1e6u, 7fd1, 1e58, 1tjy, 1me4, 1v05, 1ujp, 1mnn, 1j98, 1hqj, 1ox0, 1i1j, 1mso, 1w0n, 1t6f, 1lyq, 1i27, 1rg8, 1aho, 1pqh, 1sn9, 1usm, 1v8h, 2tps, 1gxm, 1mn8, 1r6j, 1j77, 1nwz, 1n8v, 1f94, 1rku, 1hw1, 1ouw, 1v6p, 1e29, 1fl0, 1gut, 1q35, 2bhu, 1pb7, 1c52, 1wbe, 1kng, 2aeb, 2bcm, 1k5n, 1jbe, 1ojr, 1ucr, 1jz8, 1gnl, 1kmt, 1ryo, 1pmh, 1odz, 1sfs, 1h4g, 1o4y, 1r29, 1gxu, 1vyk, 1mun, 1lqt, 1t3y, 1r5l, 1h05, 1hzt, 1tke, 1wb4, 3sil, 1od3, 1koe, 1qwn, 1kjq, 1uyl, 2bsy, 1odm, 1euw, 1oi7, 1nkd, 1g2r, 1hdh, 1n13, 1n3l, 1i40, 1vqs, 1tjx, 1kgd, 1pz4, 1nxm, 1nls, 1t8k, 1sen, 1vmg, 1mc2, 1m15, 1g5a, 1k7c, 1jub, 1en2, 1rw1, 1u7g, 1nki, 1lu4, 1sg4, 1o7q, 1cy5, 1nh0, 1vhu, 1usc, 1rwh, 1jhg, 1hbn, 1muw, 1uww, 1gu2, 1e19, 1o7i, 1q1f, 1tg0, 2nlr, 1fo8, 1c75, 1czp, 1i88, 1us5, 2a3n, 2c60, 1rl0, 1i0r, 1egw, 2lis, 1c9o, 1qre, 1n40, 1f86, 1i0v, 1vim, 1eb6, 1wdp, 1f9y, 1nz0, 2boq, 1vm9, 1hdo, 1lyv, 1ryq, 2bt9, 1bqc, 1h4a, 1tzv, 3ezm, 1i12, 1o82, 1vju,2arc, 2bmo, 2ab0, 1gkm, 1mwq, 1fm0, 3chb, 1sby, 1gwe, 1sjw, 1hz4, 1nww, 1su8, 1jm1, 1oew, 1ht6, 1kq6, 1luc, 1s9u, 1ka1, 1uuy, 2c3n, 1eaj, 1r6d, 1mf7, 1lwb, 1j3w, 1ga6, 1dzk, 1pq7, 6rlx, 1l9l, 1ucs, 1p1x, 1gwu, 1l6r, 1d7p, 1m1n, 1fg7, 1lb3, 1llm, 1q0r, 1llf, 1uv4, 1lv7, 1v70, 1ifc, 1oxx, 3lzt, 1ix9, 1rtq, 1kr7, 1qtn, 1ooh, 1es9, 1ixh, 2b3n, 1vmh, 1gwm, 1o9i, 1jo0, 1vd6, 1m40, 1kw3, 1w1h, 2b0a, 2bji, 1byi, 1f1e, 1m0k, 2bln, 1c5e, 2mcm, 1rkq, 1qj4, 1g3p, 1u7i, 2bkx, 3vub, 1mqo, 1f8e, 1eu1, 1is3, 1vbw, 1tbf, 1kwf, 1lkk, 1os6, 2aml, 1lug, 1d4t, 1ji7, 1ocy, 1n4w, 1t6u, 1g6s, 2cyg, 1vl7, 2c3v, 1vkk, 2etx, 1sau, 1uwk, 1dy5, 2fe5, 1qdd, 1oqv

C.4 Rotamer lock test and training set

C.4.1 Sequence identity 80-90

(1lfy:B, 1iwh:B), (1eap:A, 1mam:L), (1k20:B, 1i74:A), (1bx2:A, 1iea:C), (2rmp:A, 1mpp), (2dhb:B, 1hco:B), (1ncd:L, 1acy:L), (1gaw:B, 1qfy:A), (1a4k:A, 1i7z:A), (1a45, 1h4a:X), (1om3:M, 1qlr:A), (1vhw:A, 1ecp:A), (1mlc:A, 1fbi:L), (1g3q:A, 1ion:A), (1npl:A, 1jpc), (1k6p:A, 1bdq:A), (1qqp:3, 1fmd:3), (1a9q, 1ula), (1nz4:A, 1a6m), (4rub:A, 1gk8:A), (1i0a:D, 1auw:A), (1mlc:B, 1afv:K), (1b43:A, 1mc8:A), (1qh4:A, 1i0e:A), (1g3k:C, 1ned:A), (1a8m:A, 2tnf:A), (1vba:1, 2plv:1), (1llr:F, 1ltr:F), (1uam:A, 1p9p:A), (1bd2:A, 1hsa:A), (1b8d:B, 1lia:B), (1a5d:A, 1elp:A), (1dq0:A, 1mvq:A), (5gpb, 1fa9:A), (1rhi:2, 1k5m:B), (1fne:A, 1j8h:A), (1qna:B, 1ytb:A), (1f1u:A, 1q0o:A), (1g9n:G, 1g9m:G), (1rk9:A, 1s3p:A), (1k8q:B, 1hlg:A), (1agc:A, 1hhh:A), (3gtu:A, 1gtu:B), (4cln, 1osa), (2mpr:A, 1af6:B), (1lp9:L, 1nfd:A), (1mci:B, 1bjm:A), (1q0k:A, 1t6u:A), (1b15:A, 1mg5:A), (1xtc:A, 1lts:A), (1frs:A, 1zdi:A), (1lw6:E, 1mee:A), (1lt5:D, 1chp:D), (1et7:A, 1nif), (1l8f:A, 3eng), (1orq:A, 1sy6:L), (1kho:B, 1gyg:B)

C.4.2 Sequence identity 70-79

(1lpd:A, 1qi7:A), (3nul, 1cqa), (1igc:H, 1acy:H), (1qqp:1, 1fmd:1), (1h0p:A, 1cyn:A), (1aqy:A, 1hy3:B), (1l1e:B, 1kp9:B), (1ezr:C, 2mas:A), (1b0l:A, 1b7u:A), (1iwp:L, 1dio:A), (1ohz:A, 1aoh:A), (1glh, 1gbg), (1dsf:H, 1mqk:H), (1amk, 1tpe), (2bls:A, 1bls:A), (1esp, 1hyt), (1a4f:A, 2dhb:A), (1q13:A, 1ry8:A), (1mc2:A, 1ppa), (1j5o:L, 2hrp:M), (1hil:D, 1mf2:N), (1bcp:B, 1prt:C), (1gpl, 1bu8:A), (1fj1:A, 1otu:F), (1kjv:A, 1k8d:A), (1jhl:H, 1nmb:H), (1pkq:B, 1uwe:V), (1kew:B, 1bxk:A), (1jw1:A, 1b7z:A), (1cfn:A, 1rur:L), (1jqb:A, 1ykf:A), (1jk8:A, 1iak:A), (1jeb:D, 1dxt:B), (1emy, 1bvc), (1mi5:E, 1qse:E), (1f1m:B, 1ggq:A), (1qbm:L, 1ae6:L), (1h4i:A, 1lrw:A), (1fr6:A, 1kvl:A), (1m06:F, 2bpa:1), (1j96:A, 1q5m:A), (1tgk, 1tfg), (2iad:B, 1jk8:B), (1kel:H, 1cbv:H), (1hfq, 1dr3), (1cwp:C, 1js9:B), (12e8:L, 1ggi:L), (1f3d:H, 12e8:P), (1fr1:A, 1pi5:A), (1f58:L, 1fdl:L), (1tfh:A, 1a21:A), (1a0c:A, 1a0d:B), (1mpu:A, 1hq8:A), (1kt2:B, 1bx2:B), (1sy6:H, 1k4d:A), (4rub:T, 8ruc:I), (1ydf:A, 1ys9:A), (1e03:L, 1att:A), (1hbo:E, 1e6v:E), (1b3z:A, 1i1w:A), (3cbs:A, 2cbr:A), (6gst:A, 1hnb:A)

C.4.3 Sequence identity 60-69

(1ksw:A, 1qcf:A), (1vm1:A, 1nym:A), (1zib, 1pzc), (1ge2:A, 3lzt), (1nq7:A, 1s0x:A), (1p48:A, 1te6:A), (1f6t:B, 1nsq:A), (1agd:B, 1bz9:B), (1ib4:B, 1czf:A), (1dgj:A, 1vlb:A), (1nwo:A, 1cc3:A), (1lkz:A, 1m0s:A), (1tk4:A, 1s8i:A), (1mwv:A, 1itk:B), (1gil, 1tag), (1sei:A, 1an7:B), (1fbi:H, 1fj1:B), (1mlc:D, 1fj1:D), (1a9b:B, 1cd1:B), (1cob:B, 1xso:A), (1fut, 9rnt), (1j0d:A, 1rqx:C), (1ew3:A, 1gm6:A), (4ubp:C, 1ef2:A), (1prc:M, 1eys:M), (1n7g:B, 1t2a:A), (1pca, 1aye), (1gjq:B, 1gq1:B), (1h43:A, 1a8f), (1vcq:B, 1svp:A), (1dy2:A, 1dy1:A), (1kp0:B, 1chm:A), (1ubh:L, 1frf:L), (1rlg:B, 1pxw:A), (1cvz:A, 1gec:E), (1te1:B, 1h1a:A), (1h5y:B, 1ka9:F), (1bfo:B, 1uz6:W), (1okr:B, 1sd4:A), (2ae2:A, 1ae1:B), (1bfo:A, 1uz6:V), (1l2l:A, 1ua4:A), (1ppn, 1meg), (1de0:B, 1cp2:A), (1q72:H, 1f8t:H), (1m08:B, 1bxi:B), (1t45:A, 1rjb:A), (1p3j:A, 1s3g:A), (1kx5:C, 1f66:G), (1e3a:B, 1cp9:B), (1a3q:A, 1nfk:A), (2ckb:B, 1nfd:B), (1dz0:A, 1joi), (1hcb, 1hca), (1f28:A, 1hvy:A), (1d2m:A, 1d9z:A), (1kmy:A, 1eir:A), (7taa, 2aaa), (1h1h:A, 1k2a:A), (1n63:B, 1ffu:E), (1ji6:A, 1dlc), (1aom:B, 1bl9:B), (1nse:A, 1df1:B), (1bjf:A, 1g8i:B), (1edy:B, 1ayo:A), (1bd2:E, 1jck:A), (1bre:F, 1nmc:C), (1m6w:B, 1mgo:B), (1ahw:E, 1kno:F), (1fvc:C, 1ap2:A), (1a8u:A, 1hkh:A), (1vls, 2asr), (1pah, 1toh), (1j7d:B, 1jat:A), (1om4:B, 1dwv:A), (1ygh:A, 1cm0:A), (1a70, 1awd), (1k94:A, 1juo:A), (1kip:A, 1jhl:L), (1hms, 1adl), (1isn:A, 1j19:A), (1e0s:A, 1mr3:F), (1jzi:B, 1nwp:A), (1ud6:A, 1bli), (1fa2:A, 1byb), (1i3r:F, 1lnu:F), (1tyt:A, 1bzl:A), (1fvc:D, 1ar1:C), (1fsk:I, 1v7n:K), (1jnh:D, 1bfo:H), (1frf:S, 1ubh:S), (1nio:A, 1mrj), (1f6l:L, 5lve:A), (3seb, 1ste), (1onr:A, 1f05:A), (1a2s, 1cyj),

C.4.4 Sequence identity 50-59

(4tf4:A, 1g87:B), (1sac:A, 1b09:C), (1msd:A, 1mng:A), (1dsz:B, 1dsz:A), (1gow:A, 1qvb:A), (1fob:A, 1hjq:A), (1pk6:B, 1pk6:C), (1b2p:A, 1niv:A), (2atj:A, 1qgj:A), (1buv:M, 1jiz:A), (1w1z:A, 1w5m:A), (1c3a:A, 1bj3:A), (1e5m:A, 1b3n:A), (1l2j:A, 1l2i:A), (1uds:A, 1r6m:A), (1iax:B, 1b8g:A), (1kkr:B, 1kcz:A), (1ihx:C, 1k3t:B), (2rhe, 2imn), (1o73:A, 1ezk:A), (1nf3:A, 1s1c:A), (1krn, 1pkr), (1bxv:A, 1b3i:A), (1nd1:A, 1wni:A), (1jug, 1lmc), (1fx5:A, 1qos:A), (1iru:H, 1ryp:V), (1u98:A, 1xp8:A), (1ixx:A, 1v4l:A), (1ea9:D, 1gvi:A), (1rhi:3, 1nd3:C), (1i1n:A, 1r18:A), (1b3b:A, 1euz:F), (1ojo:A, 1i8q:A), (1f5v:B, 1bkj:A), (1kkl:I, 1kkm:I), (1coy, 1b8s:A), (2fus:A, 1vdk:A), (1qqw:B, 1a4e:A), (1uvq:A, 1jws:A), (1rd8:D, 5hmg:B), (1ejb:A, 1kyv:E), (1kyn:B, 3rp2:A), (1ck4:B, 1aox:A), (1hkj:A, 1syt:A), (1axn, 1ann), (1sm3:H, 1t3f:B), (1mec:1, 1tmf:1), (1mi5:D, 1qse:D), (1jsd:B, 2viu:B), (1h8l:A, 1uwy:A), (1fwb:B, 1ubp:B), (1n61:C, 1ffv:F), (1cnu:A, 1f7s:A), (1omj:A, 1sat), (1cqw:A, 1iz8:A), (1gyc:A, 1a65:A), (1a6e:A, 1a6d:B), (3tim:B, 1tph:2), (1mfm:A, 1jcv), (2er7:E, 1apt:E), (1ref:A, 1hix:B), (1dxr:H, 1eys:H), (1xr5:A, 1xr6:A), (1h6f:B, 1xbr:A), (1hl4:A, 1to5:A), (1goc, 1ril), (1lvg:A, 1ex7:A), (1e2y:G, 1qmv:A), (1n0x:M, 1ob1:A), (4eug:A, 4skn:E), (1lmh:A, 1lqy:A), (1llq:B, 1qr6:A), (1www:W, 1hcf:A), (2gd1:Q, 1gyp:A), (1p0m:A, 2dfp:A), (1bjj:A, 1kvo:A), (1uj3:B, 1qkz:H), (1wad, 2cym), (1aq0:A, 1ghs:A), (1h4l:A, 1fin:A), (1hdg:O, 1obf:O), (1ihy:A, 1ml3:A), (1msa:A, 1xd5:A), (1n4x:M, 1j05:A), (1mvm:A, 1k3v:A), (1bk1, 1xyn), (1dpf:A, 1mh1), (2not:A, 1s6b:A), (1gmm:A, 1uy3:A), (1a05:A, 1dr0:A), (1pva:A, 1a75:B), (1lbq:B, 1hrk:A), (1iof:B, 1a2z:A), (1ntz:D, 1p84:D), (1aks:B, 1gg6:C), (1ceb:A, 5hpg:A), (1iq4:A, 1mji:A), (1gnx:A, 1np2:A), (1bxb:A, 3xin:A), (1etz:L, 8fab:A), (1jb9:A, 1fnc), (1n3y:A, 1ido), (1poa, 1kvw), (1dsf:L, 1jv5:A), (1gpi:A, 1egn:A), (1e3e:A, 1hdx:A), (1pkl:A, 1aqf:B), (1ggp:B, 1abr:B), (1gjo:A, 1vr2:A), (1iwa:A, 1rsc:A)

C.4.5 Sequence identity 40-49

(1jq5:A, 1kq3:A), (1bsg, 1bue:A), (3mdd:A, 1buc:A), (1epz:A, 1rtv:A), (1upm:P, 1rbl:M), (1i8f:F, 1jbm:A), (1kr4:A, 1j2v:A), (1os8:A, 1pq8:A), (1nhk:R, 1ehw:A), (1ie0:A, 1vh2:A), (1btm:A, 1mo0:A), (2ucz, 2aak), (1hlc:A, 1qmj:A), (1i6i:A, 1goj:A), (1d8u:A, 1bin:A), (1tmo, 1eu1:A), (1pbk, 1yat), (1a6v:L, 1qfw:M), (1oau:H, 43c9:F), (1cne, 1i7p:A), (1ypr:A, 1f2k:A), (5pal, 1b8r:A), (2cel:A, 1eg1:A), (1jxn:A, 1wbl:A), (1c40:A, 1i3d:A), (1bxn:L, 1bwv:W), (1ho3:B, 4pga:A), (1ayy:C, 1apy:A), (1td5:D, 1tf1:A), (1m6d:A, 1o0e:A), (1fue:A, 1czu:A), (1kbb:A, 1l0p:A), (1psq:A, 1qxh:A), (1m1m:B, 1mzj:B), (2prd, 1faj), (1pby:A, 1jmx:A), (1o0y:A, 1mzh:A), (2qwc, 2bat), (1srd:A, 1do5:A), (1asm:B, 7aat:A), (1fvu:D, 1umr:D), (1bzy:A, 1cjb:C), (1a4s:B, 1a4z:A), (1em1:B, 1avm:A), (1eg5:B, 1p3w:A), (1l6s:B, 1gzg:B), (1j6x:B, 1inn:A), (1g0h:A, 1vdw:A), (1avu, 1tie), (1m7p:B, 1m6j:B), (1onf:A, 3grs), (1llt:A, 1icx:A), (1d0i:H, 1uqr:K), (1k44:C, 1jxv:B), (2hhe:D, 4hhb:A), (1b4p:A, 1bg5), (1eb9:A, 1y7i:B), (1e51:B, 1b4k:A), (1pvv:A, 1oth:A), (1g6i:A, 1fo2:A), (1chr:B, 1muc:A), (1mfe:L, 2rcs:L), (1s5v:A, 1o5k:A), (1nov:C, 2bbv:C), (1kmm:C, 1ady:A), (1p7c:B, 1p6x:A), (1t3q:A, 1n60:A), (1f3o:A, 1b0u:A), (1kqy:A, 1cnv), (1dxv:C, 1dxv:B), (1n4o:B, 1hzo:A), (1aj0, 1eye:A), (1naq:A, 1osc:F), (1m1t:A, 1wdk:C), (1jez:B, 1j96:B), (1aks:A, 1kdq:A), (1qhp:A, 1ciu), (1gs0:A, 1uk0:A), (1obo:A, 1ag9:A), (1pno:B, 1pt9:A), (1pee:A, 1ikj:A), (1d9q:B, 1fta:A), (1o6e:B, 1fl1:B), (1civ:A, 1b8p:A), (1arv, 1mn2), (1sc0:A, 1vh9:A), (1jd0:A, 1rj5:A), (1nh2:D, 1nvp:D), (1g0c:A, 1qi0:A), (1bwl:A, 1h62:A), (1k62:B, 1tj7:A), (2eif:A, 1bkb), (1kgz:B, 1v8g:A), (1c9w:A, 1mrq:A), (1oij:B, 1os7:A), (1j35:B, 1j35:A), (1p15:B, 1rpm:A), (1tfu:A, 1o6b:A), (1zin, 1aky), (1kcd:A, 1ia5:A), (1en6:A, 1ues:A), (2ecp:A, 1ygp:B), (43c9:H, 1bzq:K), (1j4b:A, 1cg3:A), (1qrd:A, 2qr2:B), (3erk, 1a9u), (1pwo:A, 1c1j:A), (1jeh:A, 1lpf:A), (1epx:A, 1fdj:A), (1kxt:B, 1rvf:H), (1is3:A, 1c1f:A), (1qtf:A, 1due:A), (1h2b:B, 1r37:A), (1a2d:A, 1ftp:A), (1hg0:A, 4eca:A), (1gxd:D, 1uea:B), (1m85:A, 1gwh:A), (1bla, 1qql:A), (1hm6:A, 1anw:A), (1dfo:C, 1ls3:A), (2min:A, 1mio:A),

C.4.6 Sequence identity 30-39

(1o17:D, 1khd:B), (1e6c:A, 1kag:A), (1eun:A, 1vlw:B), (1auy:A, 1qjz:A), (8dfr, 1dyr), (1f77:A, 1sxt:A), (1cgo, 1cpq), (1m06:G, 2bpa:2), (1j2y:A, 2dhq:A), (1kex:A, 1d7p:M), (1cpc:A, 1lia:A), (1bi9:D, 1euh:A), (2ans:A, 1ggl:A), (1paf:B, 2aai:A), (1lld:A, 1guy:C), (1oa4:A, 1ks4:A), (1poy:1, 1a99:A), (1ipf:A, 1pr9:A), (1eve, 1akn), (1obr, 1arl), (1ta3:B, 1n82:A), (1nol, 1hcy), (1qnn:A, 1gn4:A), (1fgi:A, 1k2p:A), (1jt3:A, 1qqk:A), (1ez4:D, 1guz:A), (1hwn:A, 1ce7:A), (1fjm:B, 1s95:B), (1aij:L, 1aij:S), (1ryp:F, 1ryp:D), (1ayy:D, 1apy:D), (1l0l:F, 1p84:G), (1p33:B, 1nfr:A), (1geq:B, 1k3u:A), (1pty, 1jln:A), (1vfs:A, 1niu:A), (1ouz:B, 1ouz:A), (1tc2:B, 1d6n:A), (1eaw:A, 1pyt:D), (1fi8:B, 1sgf:G), (1mfd:H, 1ghf:H), (1p7w:A, 1tk2:A), (1mio:D, 1qgu:B), (1kg8:A, 1e12:A), (1hrd:A, 1gtm:B), (1pd2:2, 1oe8:B), (1okt:A, 1gse:A), (1umb:D, 1ni4:D), (1yc0:A, 1kli:H), (1f4q:B, 1alv:A), (1jlr:A, 1o5o:A), (1lkl:A, 1qad:A), (1isg:A, 1lbe:A), (3sdp:B, 3mds:A), (1c8o:A, 1as4:A), (1a7v:A, 1gqa:A), (1ore:A, 1l1q:A), (1n8p:A, 1gc2:C), (1ia8:A, 1jks:A), (1j33:A, 1jh8:A), (1ng1, 1fts), (1rxk:B, 1rav:A), (1q0u:B, 1t6n:A), (1nam:A, 1h5b:A), (1e6k:A, 1nxp:A), (1g65:A, 1iru:E), (1ewc:A, 1eu4:A), (1g0o:C, 1rwb:A), (1dg5:A, 3dfr), (1f75:B, 1v7u:A), (1hp7:A, 1jrr:A), (1brt, 1a8s), (1rv9:A, 1t8h:A), (1ds7:B, 1vfr:A), (1ouw:C, 1j4u:A), (1uh7:A, 1fq6:A), (1all:A, 1on7:B), (1q7c:A, 1vl8:B), (1ey3:C, 1hzd:A), (1ls5:A, 1bim:A), (1jne:A, 1lg1:A), (1ic6:A, 1thm), (1n9w:B, 1b8a:A), (1o4s:B, 1v2f:A), (1ig8:A, 2yhx), (1c2r:A, 1lfm:A), (2hlc:A, 1trn:B), (1xqm:A, 1xa9:A), (1dys:A, 1qk0:A), (1mq7:A, 1eu5:A), (1lmw:B, 1bml:A), (1p3u:A, 1sk7:A), (1gka:B, 1gka:A), (1qbv:H, 1cvw:H), (1qxy:A, 1c24:A), (3ljr:A, 1pn9:A), (1a6z:A, 1qo3:A), (4pbg:A, 1gon:A), (2nap:A, 1aa6), (1ja9:A, 1edo:A), (1uir:B, 1mjf:A), (1e5f:B, 1ibj:A), (1pdk:A, 1l4i:A), (1fot:A, 1h1w:A), (1lqk:A, 1r9c:A), (1fj2:B, 1auo:A), (1jta:A, 1bn8:A), (1ktc:A, 1szn:A), (1iqr:A, 1qnf), (1akr, 3nll), (1npd:B, 1nvt:A), (1elg, 1dst), (1f4d:B, 1bkp:A), (1qym:A, 1ixv:A), (1rwg:A, 1j0m:A), (1g57:B, 1snn:A), (1gbl:A, 2sga), (1kfw:A, 1itx:A), (1geg:E, 1i01:E), (1qgh:A, 1ji4:A), (1k3p:B, 1aj8:A), (1i2a:A, 1mzp:A), (1cyf, 1oaf:A), (1luc:A, 1bsl:B), (1f73:D, 1fdy:A), (1j93:A, 1jph:A), (1ljt:B, 1dpt:A), (1ie9:A, 1osh:A), (1m4v:B, 1v1o:A), (1pzx:B, 1vpv:A), (2ay7:A, 1ajs:A), (1iqq:A, 1bk7:A), (3rap:S, 1zbd:A), (1dux:F, 1pue:F), (1eof:A, 3kvt), (1lme:A, 1ix1:A), (2plc, 1gym)

C.5 Representative structures used for deriving the ROTA scoring function

1x7b, 1r55, 1qs4, 1ii7, 1cul, 1fvt, 1g3d, 1g3c, 1kph, 1g49, 1r8e, 1ik3, 1ghy, 1if2, 2f67, 1m6w, 2f62, 2j0p, 1tt0, 2f64, 1lbl, 1jcm, 2b1g, 1fdj, 1nf0, 2fjk, 1ok4, 1zoa, 2eve, 2ai1, 2b3n, 1v8f, 1yad, 1jvl, 1jyv, 1jyw, 2f07, 1jz3, 1jz5, 1t4v, 1foi, 1l4f, 1ndf, 1xl8, 1fcz, 1nme, 1zd3, 1tmg, 1u0m, 1dck, 1r6w, 1qj1, 1qj6, 1ou6, 1kpg, 1wog, 2dkc, 1frz, 2o28, 1i1d, 1t5d,

2amb, 2b1z, 1u1d, 1u1e, 1u1f, 1fcx, 1xkw, 1y1z, 1yye, 1z95, 2ojg, 1tz2, 1y20, 2oem, 1z2l, 1lgw, 2afx, 2i19, 2f6t, 2abi,
2as6, 2hoc, 1t9b, 1add, 2cfu, 1pzi, 2cfz, 3gal, 2f6x, 1bky, 1t9d, 1aet, 2h7r, 1zpl, 1w8l, 1oj9, 2pcp, 2f35, 1tm3, 1scf,
2bi1, 1x2t, 1p7k, 2g8y, 2fxa, 1u2o, 2ci8, 1d7b, 1up3, 1i4f, 2hte, 1e0b, 2arv, 1ne8, 1pin, 1n5s, 2baj, 2ga4, 1gii, 1bw9,
1l5k, 2hnc, 1t9c, 2ohk, 2b0m, 2d6b, 2f8i, 1h0w, 1rs2, 2hmp, 2foe, 2g5v, 1owh, 2dvx, 2f57, 2izt, 2g5n, 2hhw, 1x78,
1aeh, 1o4q, 1aen, 2a9z, 2f7i, 2anz, 3tdt, 2fpz, 1qa0, 1u3q, 1xoi, 1vyw, 2ajc, 2oua, 1l4m, 1l4n, 2fxl, 7rnt, 1txc, 1eyn,
1ow4, 1aeo, 1n23, 2a9w, 2e4v, 1wbo, 2is7, 1ovh, 2byh, 2byi, 1jz4, 2aog, 1nn0, 2cct, 1aeq, 1pk9, 1z34, 1jz2, 1pfv, 1zai,
1umg, 3bu4, 1gmp, 2dtw, 1v5z, 1zg3, 2as2, 1i9z, 1tzk, 1rr2, 1ov5, 1aed, 1w3r, 1l5o, 1aeu, 1zp5, 1f06, 2o7n, 1ofe, 1nx4,
1gp4, 2fta, 1o57, 1w6t, 1un7, 1yc2, 1icp, 2c4b, 1bs4, 1pe1, 4tim, 1o5x, 1eqj, 7enl, 1aj2, 1c9y, 6enl, 1te2, 2aot, 2haw,
1gij, 2oh0, 2f7p, 2gj4, 2nuv, 2otf, 2c3u, 2h02, 1wnz, 2hxm, 1u3r, 2ok1, 1lc8, 1kbo, 1l4k, 2bx7, 1bhn, 2ouu, 1zrk, 2duv,
2hu6, 1exa, 1u9e, 1pu7, 1n20, 1aef, 1qw4, 1xve, 1d1y, 2cgw, 2h6b, 2b77, 1li3, 1s2g, 2cgx, 1y2c, 2fim, 1e3b, 2as4, 2bxv,
6gsp, 1goy, 1w3t, 1u1w, 2dkh, 3pcb, 2cw6, 3pce, 2ay6, 1oyo, 2g1r, 2a7p, 1xes, 1w7h, 3mag, 3pax, 3mct, 1kkr, 1i9c,
1xyc, 2hos, 1aeb, 1z3n, 2hmo, 1gt1, 2h4x, 1m7o, 1hdi, 1ejj, 1iih, 1vpe, 2cun, 1aa1, 1qhf, 2jdd, 1siw, 1cul, 1o5x, 2fme,
1vr0, 2noa, 2gm9, 2hha, 1zz3, 1zml, 1zmn, 1yc4, 1pq9, 2b1v, 2fai, 1vjy, 2h7p, 1x1j, 1x1i, 1jil, 1owd, 1f8e, 2hdr, 1yfx,
1dwv, 2qwd, 2pax, 1aeg, 1xql, 1pb9, 1yc1, 2ag6, 2ccs, 1d1x, 1s6h, 1tx7, 2bvr, 1y2d, 2aov, 2f7x, 1owz, 2f3p, 1s83, 1v5y,
2g1m, 3pcg, 1nq2, 1w1d, 1bwn, 1w2d, 1fgy, 2hds, 2cog, 1umc, 1i1m, 2fnn, 1no3, 2a8h, 1d1q, 1xkb, 2ojf, 1xn0, 2iw9,
1xn0, 2ay8, 1ps6, 1xq0, 1xpz, 1zgv, 2h4k, 2hb1, 2bpm, 1fcy, 2iiv, 2h7i, 1pl6, 1pme, 1rw8, 2cc2, 2fb3, 4req, 1lgx, 1rwq,
1rzy, 2ih9, 1xvc, 2c2w, 2i1m, 1y2e, 2eu2, 1iyb, 1c9k, 1j1g, 2ble, 1g7c, 1znx, 1jxm, 1ex7, 1qk3, 1t9s, 1mrs, 2i6a, 2c47,
2f2t, 1jhm, 1f8y, 1li6, 1m9q, 1jhp, 1v3v, 1j0d, 1d7r, 1xm6, 1xbv, 1tku, 1ohs, 1qyw, 1rxc, 1nqw, 2jav, 1oxf, 1qb6, 1z89,
2h7m, 2h7l, 1ttm, 1owe, 1tx2, 1zky, 1x76, 1c0i, 1h1r, 1y2h, 2f3q, 2iq0, 2flb, 1m9m, 2iyo, 1ko8, 1qy4, 2cxr, 2h3e, 1zsj,
1x70, 2o2u, 2h7n, 1u6q, 1c84, 1o4h, 1s63, 2bok, 1o4r, 1zaf, 1o4f, 1o4p, 1no6, 1n22, 2g6j, 1y2k, 2f3s, 1i14, 1pvs, 1foj,
2g24, 1h2u, 2flr, 1m9k, 2dbp, 2oo1, 2ovj, 2etm, 1l5m, 1l5l, 1x8b, 1i30, 2iit, 1ecv, 2gc8, 2fvc, 5rhn, 1ro9, 2h8z, 2f3u,
1xqp, 2a5b, 1c3x, 1db4, 1z11, 1yqy, 1kbq, 2b1i, 1onz, 1rsi, 2azr, 1f8d, 1egy, 1fv0, 2de7, 1l1q, 1fsg, 1a9p, 1i80, 1rrw,
2oyk, 1ik3, 1lv8, 1mr2, 1hpu, 2ga2, 1tv5, 1d3h, 2ccv, 1h1h, 2gz2, 1o0o, 1tzj, 2d06, 1x8l, 1aqu, 1q1z, 1efh, 1ka1, 1hi4,
1zrh, 1t8u, 1o0f, 1tky, 2hl1, 1yw7, 1rri, 1uj6, 1fy6, 2cxp, 1ut6, 1e3w, 1qco, 2g63, 1dry, 1jha, 1ajn, 1vru, 2i7c, 1jq3,
1wma, 2aeb, 2esl, 2d7f, 2j60, 1o8b, 1d3v, 1c2k, 1naa, 2axr, 1dku, 2eus, 2bza, 2hxc, 1m18, 2okk, 1xoe, 1pl1, 1zk1,
2ki5, 1y1m, 2gq8, 2a01, 1cea, 3kiv, 1qiq, 1vyg, 1adl, 1cvu, 2g5t, 1ukq, 2ivd, 2ivs, 1g7b, 2i7n, 1mxd, 1qjf, 2oxj, 1tjx,
2bw4, 1lqu, 2rth, 2f97, 2j46, 2o1q, 2cio, 2fnt, 2bjs, 2avs, 1r4f, 1zn7, 1wei, 1cb0, 1z8d, 1qci, 1jh8, 1m2t, 2ga4, 1s2d,
1aha, 2g5p, 1jx4, 1n5s, 4cpp, 1ho5, 2doj, 1jdv, 1n3z, 1jg3, 1pg2, 2jc9, 1yi4, 1v8b, 1lii, 1dad, 2oxc, 1xnj, 1b62, 1gzf,
2c9o, 1z59, 1oxu, 1sdm, 2iyz, 1eq2, 1mqw, 1xnj, 1v47, 1o9u, 1yb1, 1x8j, 1nfq, 2c2n, 2c2n, 2b4r, 1re0, 1uum, 2bgr,
1iuc, 1mt1, 1y5i, 1ejn, 2a0z, 2nzt, 1v2b, 2f2e, 1s5m, 1hor, 1mos, 2og7, 2ay1, 2f92, 1fpl, 2g7q, 2afw, 1wd4, 1ki6, 1fd7,
2hj9, 2o3z, 1h8s, 1nx9, 2g5j, 1oxr, 1tgm, 1k4g, 2fwj, 1zr8, 2fdi, 2cjh, 2j9e, 1e5i, 2gp5, 2fcu, 1yba, 1cw4, 1mzf, 1oij,
2air, 1ung, 1w1v, 1rpj, 1r6n, 1az1, 1afe, 1m5e, 1tx8, 1ama, 1dlg, 1tng, 1sja, 1jac, 1ceb, 1jir, 1fa9, 2hbl, 2j91, 1eyk,
1o97, 2gm3, 1wxi, 1zn9, 1ua4, 2j9d, 1my2, 1f5l, 1aev, 2cnq, 1m9n, 1zq8, 1lo3, 2hmn, 1h62, 1gt1, 1coy, 1ifs, 1j1r, 2bjm,
1lwv, 1ppa, 1hj9, 1aee, 1sv3, 2b96, 1qpc, 1jpa, 2dxd, 1zl2, 2bfg, 1r5g, 1r5h, 1r58, 1wyv, 2f7q, 1lho, 1lhn, 1zq6, 2i57,
1dku, 5fit, 1av5, 1dmj, 1dmk, 1tpp, 2dap, 1gg6, 1k4h, 1vhz, 1v8y, 1e2i, 1enu, 1s39, 1p5z, 1mmz, 1abe, 2arc, 1abe,
1h69, 1nq7, 1e2i, 1vag, 1dcn, 1j20, 1pqp, 2gz3, 2aa7, 1f9g, 1oaf, 1e71, 1eup, 1qyx, 1xf0, 1nas, 1n4f, 1oc1, 1n0t, 2ha5,
2tmk, 1w2h, 1e9d, 1tc0, 1n48, 1u5v, 1obg, 1d1w, 1wxi, 1q3w, 1x2e, 1dzt, 5prc, 2i03, 1dil, 1lxc, 1tuf, 1i7g, 2icq, 1j2g,
1vay, 1kti, 1lwx, 1fen, 2j7h, 1jd0, 1yda, 1oar, 2bxk, 2foq, 2fos, 2h15, 1p06, 1rej, 1tou, 1p02, 2ccr, 1p03, 2foy, 2b4b,
2hfe, 1t67, 2bzg, 1tb7, 1d6f, 2c1x, 1rv6, 2je8, 2ooq, 2awh, 1g9a, 1elf, 1wda, 1n2o, 1tio, 1xx4, 1zvw, 2anw, 1f5k, 1c5o,
2cht, 1u1c, 1g27, 1ix1, 1ws1, 1lqy, 1t48, 1v2n, 1u1g, 1d8m, 1s9j, 1lo0, 1qyg, 1p0p, 1ki0, 1v0j, 2jc5, 1ltm, 1m66, 2htu,
1l7g, 1n8v, 2glp, 1ovp, 2fa1, 1ybq, 1n2v, 2bpq, 2bmv, 1v16, 1lpu, 2ayw, 2ast, 2aiq, 1w80, 1wri, 2gnn, 1gql, 1jjt, 1f8s,
2gvq, 2jb3, 1zfk, 1oon, 1v2u, 1jbu, 1cc8, 1eax, 1lo6, 1zhr, 2glp, 1nkz, 2pka, 1rtf, 1lfo, 1dtl, 1ybk, 1xry, 2hpt, 1sw2,
1rcg, 1wwj, 2b4l, 1r9l, 1gyx, 1w9e, 1s8f, 1ukb, 1rum, 1ree, 1urm, 2f7a, 1kqb, 1i7q, 1pwl, 2f48, 4std, 1e7y, 2nzt, 2cir,
1ofl, 2ipn, 2j0y, 2cn3, 2b3b, 2cdb, 1l1y, 1sz2, 2bwc, 1taq, 1l9n, 1eys, 1nwg, 2afu, 4nos, 1mmt, 2dtt, 1lcz, 1sxk, 1q65,
2hi4, 2h88, 1ppj, 1m00, 1xvb, 2ax9, 1hsr, 2atj, 2boy, 2agv, 1dmi, 1uu9, 1zyj, 1u4s, 2amv, 2bd0, 2a0s, 1sep, 1dr4,
1b66, 1r1h, 1yv3, 2a4x, 1tqw, 1mzn, 2bgr, 2j0y, 1w1x, 1h1m, 1nb5, 2gh0, 1w7c, 2dtx, 1qkq, 2gud, 1qo0, 1g55, 1zsw,
1qtz, 1x8i, 1xu3, 1t0s, 1dqx, 1lor, 2rma, 1m5b, 1wuy, 1wv0, 1wv1, 1u3u, 2ic8, 1ehk, 1j4n, 1ymg, 1rxk, 2fhn, 8kme,
1p05, 1pq9, 223l, 2a31, 1p01, 1tcb, 1u7g, 2jaf, 2fsm, 1xez, 1d2m, 2cmc, 1yc9, 1nf9, 1fx8, 1dqe, 2hd6, 1op3, 1s2i, 1n24,
1lgt, 1lkd, 1qhi, 1aax, 1yeg, 1d4p, 2de4, 1kwc, 1kqu, 1bso, 1mq0, 1d3g, 1uuo, 1lww, 1m5c, 2izr, 1zgy, 2izs, 1rl4, 1gxz,
1o27, 2af6, 2bhe, 1k3t, 1yqs, 1g4o, 2gwc, 1i76, 2bsm, 2c1s, 19gs, 1tw6, 2ob3, 2i3a, 2c5a, 1xa3, 2bib, 1vcl, 1rj4, 1r64,
2hvr, 1ki4, 1wnb, 2c4i, 1swr, 1hxd, 2dxt, 1xny, 1vqn, 2f01, 1sux, 1vlf, 1tzp, 1vio, 1bk9, 2hw8, 2bab, 1lol, 1z2u, 2h90,
1xh5, 2f9p, 2cl5, 1uk7, 1ugp, 2cz0, 1cp6, 1d7j, 1rcv, 1rd9, 1rdp, 1rf2, 2euq, 1ki8, 2eut, 2euu, 1mqh, 1h08, 1h08, 1jje,
1k06, 182l, 1ryc, 1l5f, 1dzm, 1okk, 1dzp, 1gt5, 1wht, 2a3i, 2bhi, 1w6j, 1xji, 1h02, 2bxa, 2c90, 2hdq, 2h92, 1oik, 2c8z,
2aov, 1tzm, 1jvu, 2gqs, 2c8y, 1rpf, 2c93, 2e4w, 1iv4, 1xrj, 2ix0, 1lp6, 2ukd, 1h47, 2ex1, 1w77, 1h7f, 1prn, 1hxx, 2gr7,
1qj8, 1ump, 1nqe, 2o4v, 1qjp, 2hcg, 2bt4, 2oz7, 1noo, 1t87, 6cpp, 2can, 1ir2, 2buy, 1xep, 1kdr, 2shp, 1szo, 2bzs, 1idt,
1ztz, 1ivr, 2abj, 20gs, 2fbw, 1p4j, 1lt8, 1z3t, 1tvp, 2b1r, 1qi0, 3eng, 1g0c, 2dwi, 2o7i, 2ovw, 21gs, 1cxv, 1t1s, 1c7t,
1zu0, 1tw5, 2bs7, 1pzo, 2cbx, 1u3t, 2d3u, 2f7a, 1h48, 2gnu, 1ffu, 2az3, 2cmk, 2oyn, 1h7h, 1orr, 1xjn, 1uf5, 1uf7, 1qye,
2b7o, 1nnk, 6prc, 1j0n, 2cix, 1oq5, 1fj8, 2gh6, 1a3l, 1l5q, 2a3b, 1g20, 2afh, 1jdj, 2cbv, 1p4g, 1rca, 2ag2, 3pch, 2hrc,
1v55, 1tw4, 1s9q, 1ee2, 2azy, 7kme, 2ha0, 1ahi, 1w1t, 1h8g, 2ha3, 2fy3, 1tio, 1xoz, 1wcc, 1itu, 3gch, 1t6j, 1xlx, 2nz2,
1h70, 2c6z, 1kp3, 1gi7, 1z1e, 1mk0, 1gcz, 2c7p, 1cxu, 1tm3, 1re0, 2bs3, 2fqg, 1vj5, 1pxi, 1pxj, 2c5p, 1pxp, 2a0c, 2dcn,
1y2g, 1qz0, 1g20, 1lev, 1cla, 1usq, 1cet, 1lri, 1n83, 1zhy, 1kvl, 2ay7, 2hgy, 1m9j, 1m78, 1jgt, 1e1v, 1e3z, 1i6x, 1q5o,
2ouy, 1cx4, 1ykd, 1lpc, 1z5u, 1vp6, 1chm, 1uou, 2of0, 1r1n, 1a26, 1r1n, 1r1n, 1e8d, 1t2b, 4sli, 1u9n, 1n7x, 1pd9, 2jfk,

2c27, 1jkj, 2deb, 1q72, 1ly3, 1dxy, 1umd, 2hdk, 1gfy, 1hbo, 2c3c, 1ly4, 1z10, 2h90, 2aw1, 1j3j, 2blc, 1utt, 1c8k, 1rnc, 2a3y, 1i7g, 1gr3, 1puc, 1xu7, 1xji, 1syh, 1suo, 2aou, 2c1b, 2g6x, 1z1p, 1fu7, 1myw, 2dd7, 1fu8, 1b4d, 1xag, 1nr5, 1fk0, 1kzl, 1v7z, 1huy, 2std, 1uz9, 1yjf, 1yj2, 1g55, 2ae7, 1ktb, 1m7g, 1k2e, 1kbn, 2a5i, 2ev6, 1ir2, 1kae, 1tow, 1pa9, 1elu, 2ah8, 2evo, 2bt0, 2brc, 2c68, 1y8y, 2c69, 1aln, 2ar8, 2an3, 2fr6, 2oh7, 1tug, 1i52, 1ueu, 2im0, 1coz, 2a3h, 2cbu, 1eqc, 1n62, 8kme, 2g16, 2erv, 2ay2, 1n0x, 1y1j, 2a26, 2ix9, 2flk, 2dd7, 1rzl, 2bj0, 1gvq, 1b2l, 2d0q, 1lbc, 2evd, 2boz, 1xji, 2h4t, 1gzq, 2j1n, 1xji, 1ian, 2hei, 1w19, 2fdu, 2jaj, 2b53, 2fdv, 2bfn, 2gz7, 2fdw, 1w7r, 1w7r, 1w7r, 1w7r, 1w7r, 1w7r, 1w7r, 1w7r, 1tkd, 2ihm, 1xsn, 1f3f, 2fdy, 1z4q, 1w7r, 1w7r, 1w7r, 1w7r, 1w7r, 1w7r, 1w7r, 1w7r, 1w7r, 1w7r, 1f3f, 2ioc, 2pri, 1rm0, 1rs9, 1yrx, 2aqd, 7nse, 3dap, 1bs1, 1skr, 2imw, 1qsy, 1d1a, 1p4n, 1czq, 1iki, 1c0p, 2cex, 1ms1, 1wcq, 1z4v, 2qwc, 2f25, 1v3d, 1w0o, 2sim, 1sli, 2b5s, 1tjj, 2bz3, 1hzp, 1e7f, 1d1b, 1czq, 1bg0, 1zea, 7jdw, 1ov6, 1z57, 1u4d, 2c3j, 1gqg, 1g5f, 1rnd, 1b5e, 1nje, 1jw7, 2dpi, 2hhu, 2jaq, 1pkk, 2aq4, 1s10, 1d7s, 1tk0, 3ktq, 1s97, 2hvh, 2fmp, 1f57, 1rmy, 1p61, 1fm8, 1dyj, 1hxk, 1vq2, 1jsl, 2f6m, 1d3g, 1aob, 1s9f, 1znl, 1z8q, 1nfs, 1y2b, 1tr7, 1ga8, 1ek8, 183l, 1lk7, 2cxo, 1tz8, 1s9p, 3erd, 1lgh, 2bet, 2bro, 1mg0, 1jep, 1fm7, 2brh, 1uhv, 2brg, 2brm, 2h4z, 1jki, 1r3j, 2gvj, 1w7r, 1w7r, 1w7r, 1w7r, 1w7r, 1w7r, 1w7r, 1w7r, 2o08, 1czq, 1qur, 1zuw, 2asv, 1b74, 1zea, 2fc0, 1del, 1z4p, 1tlc, 2iry, 2gcg, 2bo6, 2fzh, 1gp6, 2fzj, 2fzi, 1ykl, 1n8q, 1b4u, 1phh, 2bsl, 2o7d, 1cs1, 1mxh, 1rf7, 5p2p, 1t5f, 1czq, 2ez7, 1zea, 2hzy, 1we2, 1gtz, 1e00, 1gt3, 1fm4, 1al8, 1pax, 1t63, 1kdm, 1dht, 3pcn, 1q0c, 1ctt, 2gvv, 1h83, 1m18, 2b6d, 1nr6, 1n3i, 1rsz, 1ez2, 1zb6, 2ahj, 1kkh, 1tgj, 2jen, 1h0a, 2j59, 2jc5, 1bfu, 1kxg, 1wyk, 1t4t, 1pbq, 1wbe, 1gxs, 1f91, 2chl, 2j51, 2ffr, 1czq, 1rcq, 2ats, 2ick, 1yhl, 1sri, 4gch, 1d0s, 1vrq, 1el5, 1cmp, 1fo2, 1g6i, 1kre, 2tod, 1c1e, 1c2g, 1jmb, 2avq, 2rmb, 2gsm, 1qmg, 1tp7, 2b14, 1gvo, 1a82, 1ftl, 1wz1, 1h87, 1doe, 1kdt, 1lke, 1byd, 2osu, 1xge, 1zt5, 1tw1, 1g29, 2g56, 2gnv, 1v19, 1t56, 2c0n, 2fwv, 1kmq, 1p6k, 1rs6, 1p6m, 1nd0, 1p6j, 1q91, 1a4q, 1qiw, 2fmz, 1tkh, 2dns, 1wxi, 1n5l, 2o1s, 1pvf, 2f7f, 1rqi, 1zpd, 1f3e, 1zly, 2c29, 1gp5, 1qrd, 1q4w, 2asc, 1zjp, 2a8u, 2asc, 1p4f, 1q92, 2g5o, 1daf, 1zea, 1pb8, 1hg1, 1d7i, 1rqi, 2c6i, 2c6k, 2c6l, 2c6m, 1dam, 1ooq, 2izx, 1w80, 2c2z, 2aaw, 1h16, 1mvs, 1ft6, 1qmh, 1di8, 1czq, 1tkf, 2olw, 1nyt, 2am9, 1e42, 1ywm, 2g8x, 2cko, 1o9j, 1n1c, 1dq8, 1w19, 2h00, 2bgc, 1w4r, 1w19, 1lnm, 1of6, 1zea, 2fvk, 1q5h, 1duc, 1dud, 1sjn, 2fms, 1rnj, 1oe5, 1tdu, 2c53, 1f7q, 1xs1, 1smc, 1syl, 2nom, 1pkj, 1zea, 2fli, 2bjf, 1os6, 1e3v, 1pjx, 2c2n, 1ecq, 1m5w, 1q0q, 1on3, 2f13, 2h55, 2d2g, 1zq5, 1w8m, 2j7y, 2hfk, 1jdb, 1iri, 1fwt, 1rzm, 1ngs, 1y5x, 1pu8, 2hp1, 2gss, 1j12, 1p6c, 2ajz, 2fkk, 1kcz, 1z5y, 1gm7, 2atb, 1vqz, 2cjt, 2onf, 2fi3, 2ftz, 2ack, 1zlq, 2axn, 1nnf, 1p6b, 1jp3, 2f32, 1s84, 2a5i, 2bem, 1m33, 1hx0, 1jgm, 1jnq, 2hxt, 1zs0, 1ppw, 1hmr, 1zot, 1qd9, 1o9l, 1kdg, 1w1s, 1c1e, 2f6y, 2fzg, 2dio, 1jxy, 1alz, 1alz, 1alz, 1alz, 1alz, 1alz, 1alz, 1alz, 1alz, 2fzc, 1m0q, 2faf, 1oe5, 2hhf, 1gyx, 2c7p, 1s2e, 2j3u, 1qst, 1z6g, 1vse, 1j11, 1zo9, 1qxo, 1dim, 1cqs, 1qjg, 2aib, 1zhz, 2h0y, 1o5b, 1c5q, 1c5n, 1x8v, 1lhw, 1ohp, 2d06, 1aqu, 1gwr, 1lhu, 1e6w, 1iol, 1jgl, 1c5s, 1byz, 1byz, 3al1, 3al1, 2g50, 1mxg, 2d1g, 2hcy, 1sby, 1rjw, 1p2s, 2fog, 1yl1, 2ha5, 1erb, 2hx5, 2d1g, 2gno, 2fd6, 1qlu, 1r3j, 1znk, 1yce, 2hdu, 2cm8, 2clx, 2df8, 2g72, 1g52, 1w8r, 2gz8, 1ofe, 2bs3, 1l5j, 1siw, 2h88, 2ivf, 1gao, 1g53, 1eyk, 2owz, 1uxr, 1t10, 4pfk, 1lby, 2axn, 2bpl, 2bkx, 2g8n, 1h0r, 1gu1, 1v1j, 1h0s, 5yas, 1y9d, 1n5w, 1x0p, 2hmb, 1tnh, 2b9a, 1lbz, 1a5z, 1w8s, 1pfk, 2giu, 1if4, 1if5, 1if6, 2evm, 2evc, 1ury, 1vzh, 7abp, 1h82, 7abp, 2aac, 1xnz, 1lqp, 1yrq, 1e8g, 1hfe, 1nmx, 2fyu, 1frp, 2avk, 1fel, 1gw2, 1uwc, 1jt2, 1kyz, 1gkl, 2eu3, 1g54, 2hs3, 3pcf, 1gyy, 1mfi, 1y1d, 2ax7, 1hxc, 2agt, 2d3z, 1tn6, 1xlz, 1zvx, 1fuy, 1x97, 1x98, 1kgj, 1thc, 1c0l, 1s69, 2d4v, 2bp1, 2h4v, 2bhb, 2ev6, 1yta, 1c5y, 2o66, 1zb7, 1s2c, 1ocb, 1ovj, 1r9o, 1x9q, 1n0s, 1t66, 2d09, 1zdw, 1k9s, 1k9s, 1tc1, 1jdz, 1a69, 1sd1, 1ifu, 1nc3, 1r78, 1tg6, 1j1s, 1ahb, 1uxh, 1qlb, 2aa9, 2izb, 1r4p, 1l0l, 1e3d, 2c1p, 2gah, 2gf3, 2bk3, 1ab8, 1vif, 1dyi, 1q0h, 1x9h, 1fqo, 2gc2, 1oz1, 1xvd, 1kzo, 1v7u, 1t0a, 1x08, 1rak, 1ndv, 1ndy, 2hk1, 1af6, 2dd4, 2fit, 1zx5, 1y9g, 2ieg, 1ff2, 1yrq, 1g8j, 1g20, 1ff2, 1c97, 1b25, 1ti4, 1h7w, 1kek, 2fjd, 1qlb, 1r27, 1g45, 2fhr, 1e3d, 2nw9, 1pzp, 1g4t, 2g71, 2iwg, 2ha2, 1k12, 2hox, 1u8e, 1m2t, 3kmb, 1dwh, 1iub, 1fe8, 2c4f, 2j1s, 1ofz, 1ksu, 1gz3, 1qco, 1z9y, 1mqi, 1u3w, 1v97, 2fkm, 1o03, 2i80, 1l5v, 1uxt, 3gpb, 1nt4, 1h5r, 1p5d, 2qwi, 1bji, 2ouq, 2g9r, 2qwj, 1e70, 2igo, 2qwk, 2ht8, 2okg, 1nqo, 1uxu, 6tim, 1fdj, 1of8, 1wbj, 1ixo, 1m7p, 1y38, 1z82, 1ofc, 1y0b, 1e3z, 1mmy, 1gpy, 1gz5, 1u8x, 1p5g, 2acq, 1cza, 1mor, 1up7, 1u0f, 2j6h, 1x9i, 1u0f, 1ki2, 1lt6, 1t25, 2c4w, 1g8z, 1uas, 1rvz, 1axz, 1ns0, 1gca, 1isz, 1tlg, 2cn3, 1ugw, 1men, 1ez1, 1c3e, 2cgr, 1efi, 2dxi, 2f18, 2f1a, 2f1b, 1gbn, 1t26, 2gsq, 2cbi, 2coi, 1rtk, 1bgg, 2eud, 1mqq, 2fuq, 1k9e, 1h41, 1l8n, 1o7a, 1hna, 2fe4, 1s8f, 1t2a, 1zny, 1mky, 1mre, 1re0, 1q21, 1tpz, 1z22, 1r82, 1fr8, 2j47, 1x7r, 1p62, 25c8, 2bcg, 1vg0, 4gpb, 6nse, 1tqu, 2j7d, 2j7e, 2j7f, 2j7g, 2ceq, 2ces, 1w1p, 8est, 1z4o, 1ftq, 1hlf, 1ftw, 1xc7, 1fty, 1fu4, 1oh4, 8abp, 2j3u, 1axz, 2aj4, 1w2t, 8abp, 1gx0, 1ukq, 1pum, 2j5z, 1t0o, 2by0, 1fa9, 2i5p, 1j0z, 1kme, 1af6, 2aep, 1ua4, 4tf4, 1woq, 1ukq, 1gg8, 2rth, 1xym, 1moq, 2nz4, 1un7, 1ec8, 1ggn, 1n5m, 1ih7, 1je1, 2sar, 2an9, 1odj, 2fqx, 2qwe, 2a8g, 2dxf, 1s17, 1dx6, 1u98, 1jqx, 1t1v, 1w9b, 2bjf, 1gdq, 1wbe, 1yli, 1oh4, 2j47, 1e72, 1u30, 1uwu, 2j78, 1s1d, 1bjv, 2aay, 2ggd, 1rf6, 5gpb, 1zcw, 1w55, 2amt, 2dvu, 1zly, 2j1p, 1o1r, 1k5s, 4kmb, 1fro, 1yzx, 1xw6, 1jlv, 1tdi, 1px6, 1u3i, 2cz2, 2fls, 1v40, 1y1a, 2hgs, 2c4j, 1zb6, 2caq, 2j62, 1uwt, 2j79, 1kcd, 1m99, 1v2a, 1n2a, 1ev4, 1glp, 6gsy, 5gss, 2gdr, 1gsa, 1ss4, 1k0d, 1qh5, 1tu7, 1oyj, 1pd2, 1gsu, 2aaw, 2glr, 2c3q, 1k3y, 1bh5, 2c80, 1m9a, 1tu8, 1pn9, 1jz6, 1fxu, 1iyd, 1xey, 1b4n, 1xt4, 1it7, 1sql, 1d6a, 2i9u, 1vmk, 1xe7, 1a95, 2izk, 2uv2, 1aoe, 1ws4, 1hkd, 1h82, 2fqt, 2fbz, 1tg2, 1dww, 1n2n, 1pcw, 2fhj, 1q8u, 2fet, 1zw9, 2fwz, 2h7j, 1v48, 1ai4, 1sre, 1y93, 1naw, 2h52, 1d5x, 5nse, 1m7z, 2gq9, 1oyb, 1eb9, 1z42, 1jgu, 1ben, 2toh, 2g6i, 2fc1, 1e92, 1dcp, 1ltz, 1mlw, 2bf7, 1e4n, 1p28, 1p28, 1hyo, 1zhw, 1zhx, 2o7b, 2h3w, 1g20, 1w03, 1bxg, 1v2f, 1toj, 1ay8, 1m0n, 1zht, 1q8a, 1jqw, 2ci5, 1xdj, 1lon, 1cib, 1p9b, 1zmj, 1qnf, 1ljt, 1cru, 1lic, 1oit, 1oiq, 1c8v, 1zne, 1ozh, 1zng, 2a4t, 1pkz, 1pw1, 1a0q, 2j58, 1gmd, 2gsk, 1cwq, 2evs, 1ci1, 2tio, 1v7c, 2ny0, 1x1v, 1in4, 1nhz, 1p2v, 1qud, 2hy6, 1iby, 1m8t, 2ior, 1c9d, 1o7t, 1o7t, 1qbq, 2erz, 2ax6, 1so2, 1xfg, 1bv3, 2dzb, 1rb0, 1u1x, 1y9t, 1q0n, 2dhn, 1nbu, 1ru2, 2gyu, 1yve, 1e7b, 1xz1, 2g6p, 2nq6, 2nq7, 1mfv, 1dm2, 1to9, 2f2g, 1zga, 1f73, 1fp2, 2g70, 1d6v, 2hoz, 2b5s, 1a9q, 1tjp, 1u3v, 1nww, 1yns, 1o4o, 1e2m, 2dsa, 1k70, 1p6o, 2cvd, 2ddq, 1dm7, 1uu1, 1h72, 2br6, 1jqd, 1u18, 1qft, 1kar, 1avn, 1kae, 2fpu, 1cw2, 2bgi, 1u19, 2c3y, 1v5g, 2ezu, 2f12, 1yf6, 2bf3, 2hi2, 1rp0, 1u19, 1kmo, 1lgh, 1axr, 1gpn, 1gpk, 1e66, 1mqj, 1mv9, 1fdq, 2g7z, 2byo, 2ddh, 2fjv, 1xan, 1d6y, 2fqo, 1zlt, 1wzb, 1ym8, 1xnn, 2ou3, 2p0d, 1btn, 1mai, 1h0a, 1oqn, 1n4k, 1u29, 1w2c, 1z2p, 2p0h, 184l, 1z2o, 1r35, 2c1a, 2jdo,

1eko, 1el3, 1z62, 2oyf, 1k3u, 1k8z, 1k7f, 1vcj, 1zkl, 2hd1, 1zkn, 1rkp, 1i7e, 1xz3, 1c97, 1itw, 1cw1, 1xkd, 1yb7, 1g67,
2fwp, 1lwd, 1w7f, 1w8g, 1xg4, 1ki7, 1t40, 1oxl, 8a3h, 2oyl, 1z3w, 1x38, 1x39, 2j7c, 1gth, 1aek, 1mfp, 1mrd, 2c3l, 1jak,
1now, 1uz1, 1uz4, 2g9v, 1oif, 2nsx, 1g6c, 1a5b, 1a53, 1zxc, 3pci, 1lzz, 1lr8, 2ati, 1oxo, 1sff, 1dae, 1i7q, 1ed6, 1kt8,
1t1r, 1aky, 1oau, 1aes, 1w7c, 1jzf, 1jzg, 1dqp, 1b8n, 2ff1, 1rt9, 1nw4, 1g2o, 1nc9, 1xkx, 2dv8, 2bxk, 2alt, 2dm6, 1s2a,
1p9b, 2bzn, 1meh, 2prj, 1zfj, 1qk4, 1cib, 1yfz, 1p19, 1ce8, 1z6d, 1i80, 1m18, 1dqn, 1uma, 1az8, 1ftx, 1b8y, 1o7n, 185l,
1l4h, 1om5, 1uf8, 1kyy, 2f59, 1rvv, 1ejb, 1i90, 1bkc, 3pcl, 1i91, 1v0o, 1g0i, 1i9l, 1i9m, 7std, 1i9n, 1i9o, 1hc0, 1i9p,
1i9q, 1rw1, 1rl9, 1vrp, 1ahf, 2oli, 2ay5, 2bxh, 1xbu, 1qje, 2oht, 1e06, 1yvx, 1g0h, 1lbx, 1x07, 1lph, 1foh, 1jhx, 1xu5,
1li2, 2as3, 1fjw, 2b32, 1beu, 1wxj, 1rqj, 1zw5, 1jyx, 1na3, 1ed4, 1qy2, 1ydt, 1ydr, 1yds, 1s1j, 1cjb, 1nx3, 1iup, 1nf8,
2bk5, 9nse, 2bvd, 1w6f, 1joc, 1unq, 1upr, 1b55, 1fao, 1nse, 1mjt, 1gte, 1uk9, 2d5y, 1nlu, 1mqg, 1unh, 1q41, 1opm,
1wq3, 1sdw, 2c3i, 2j90, 2b7a, 2g3f, 2jap, 1pzk, 2jam, 2c97, 1ngx, 2hzi, 1yvz, 1zoe, 1zog, 1zoh, 1ftk, 1dj9, 2f5v, 1v1a,
1q9v, 1map, 1jdf, 1qw8, 1qw9, 1fo3, 1krf, 1ps3, 1o68, 1sr9, 1kta, 2bre, 2c1z, 1h1m, 1gqh, 1vb3, 1mo9, 2cfc, 1m3u,
1ycl, 2j5s, 1way, 1wbg, 1zg9, 1wbn, 1w84, 2g21, 2hvk, 2d5x, 1rdt, 1zg8, 1tj1, 1c0k, 2fn7, 2imp, 1l3l, 1jz8, 2dyx, 1j8v,
1mid, 1rdw, 1dll, 1o7o, 1zj0, 1is3, 1it0, 1ms9, 1w6o, 1v00, 1z3v, 1puu, 1s8g, 1ma0, 1w8o, 1zyx, 1d7u, 1yf6, 2fkw,
2irv, 1xkw, 2iwv, 2gsk, 2gvm, 1yiv, 1f7s, 1thq, 1okc, 1jsr, 2i17, 1z4a, 2hxu, 2ccc, 2cgl, 2eun, 2eup, 2euo, 1q6o, 2ihq,
1e6x, 2e40, 2bl2, 1m0k, 1wbv, 1wbw, 1yxv, 1yxx, 2gtm, 2gtn, 1kms, 1kmv, 1r33, 1r34, 1c1x, 2ch1, 1nqv, 1fk6, 2byo,
1wax, 2bys, 1vqn, 2iqd, 1dp2, 2c8m, 7gch, 1ke5, 1ke6, 1ke7, 1ke8, 1ke9, 1i05, 1wap, 2no9, 1s4m, 2cc7, 1he5, 1xbz,
1gw9, 1q6q, 1e7v, 1ov7, 2ewg, 1b42, 1qxw, 2fue, 1pcj, 2iwx, 1qxy, 1qye, 1zaj, 1qxz, 2h23, 1syo, 1m6p, 1pcm, 2gc3,
2f2u, 1q8w, 1ej4, 1x92, 1tdg, 2c27, 1qnr, 1yaa, 2cst, 1b4x, 1k7h, 1aog, 4kmb, 1dq8, 5mdh, 2bex, 1ua3, 1obb, 1cxi,
1qhp, 3csc, 2gjp, 1nl5, 1wdr, 2f5t, 1k1y, 1x1v, 1kuj, 2gn3, 1s4p, 2e22, 2gnd, 1orv, 2j3u, 2aep, 1kza, 2hox, 1y9m, 1dwh,
1xoe, 1s38, 2bgm, 1jyq, 2dn1, 1dr7, 1gz8, 1hy7, 2b6o, 1on9, 1dd6, 1jt1, 1m2x, 2fu8, 1l4g, 1b5q, 2vp3, 1pr2, 1w3v,
1pjx, 1y9q, 1kq0, 1tkj, 1nv8, 2g8e, 1wrm, 2dfp, 22gs, 1tw1, 2hv8, 1ma3, 2o3b, 1zs0, 1nww, 1r31, 1y0p, 1c22, 1pfw,
1kww, 1rdj, 1rrx, 2bt9, 1jxn, 1rdi, 4dcg, 1q8u, 1afa, 1i7c, 2ao3, 1n3o, 1gic, 1loa, 1n4q, 1f4x, 2f47, 1ix1, 1srg, 1oyf,
2iuf, 2g8t, 1ami, 1tlm, 1tom, 1y6o, 1o9p, 2fah, 1o4m, 1nu4, 2apw, 1s1f, 1q44, 2fw0, 1ys4, 1s7f, 2esl, 2hg8, 2j0p, 1sc3,
2bw4, 1tug, 2b4u, 2c95, 1ufy, 2h89, 1t9f, 2ftw, 1o9o, 2img, 1xuu, 1pj4, 1amz, 2dfd, 1fup, 2gq3, 1wwj, 2fgq, 2g76,
2g9n, 2etd, 2h2j, 1yik, 1yil, 1kwu, 2d7f, 1lob, 1msa, 2g93, 1xuz, 2gvc, 1kqr, 2oym, 2alw, 1yb6, 1ai5, 1d0x, 1m2q,
1m2r, 1yrq, 2hk6, 1qgu, 2g2s, 1qu2, 1lng, 1ng1, 1xg0, 2hk6, 1nwl, 2bj3, 1r03, 1qgu, 2bhb, 1i9s, 1dci, 1meh, 1srh,
1q0y, 2fu9, 1rtw, 3mth, 1p3e, 1swu, 2esp, 2ewb, 1mkp, 2c7w, 2bja, 1hx6, 1tdg, 1kwg, 1e12, 2guf, 1p7p, 1c23, 1aem,
1pfu, 1c24, 1m0o, 1kji, 2f2h, 2eve, 1l9l, 1vkp, 2axi, 1f07, 2ay3, 1zh9, 3std, 1m79, 1m7a, 2bja, 2bhb, 2c21, 2j46, 2bfe,
1oad, 1v4s, 1t7l, 6std, 2o3l, 1vje, 1we5, 2gah, 2otm, 1yis, 2nr5, 1f8g, 2ftr, 1p2f, 1el8, 2j8r, 2f7o, 2ca3, 1dxr, 1eg2,
2o05, 1cg6, 2hte, 1jdt, 2a8y, 1z5o, 2ipx, 1srf, 1el7, 1fo4, 2a9a, 1xdy, 1el9, 1nc1, 1sd2, 1q1g, 1zzq, 1m2w, 1k27, 1y6r,
1zos, 2aa0, 1pr4, 2bii, 2c5y, 2uue, 1r4s, 1lod, 2rma, 1dlr, 2cb8, 2o63, 1gym, 2ato, 2fo0, 1p42, 1hk4, 1fk2, 2ag2, 1ted,
1pzl, 1cdk, 1icm, 1hbk, 1jdj, 2o64, 1pvn, 1tkb, 1oi9, 1nhb, 1hkn, 1h3m, 2fky, 2g8r, 2ot1, 1oiy, 186l, 1vyz, 1w8c, 1oiu,
1nx8, 1tv5, 1srj, 1nzx, 1kyq, 1v59, 1bw9, 1nm5, 1iso, 1ju9, 1wze, 1mjt, 1w9b, 2hha, 1twx, 1o7a, 2j3u, 1dx6, 2i5y,
2nt1, 1o9w, 1ksi, 2bja, 1tzm, 1zh0, 1qyw, 1usf, 1eyq, 1cgk, 2brt, 1x88, 1f74, 2j4g, 1h84, 1r5r, 1em6, 1s61, 2c0u, 2bmq,
1zd3, 2h4j, 2od9, 1r15, 2otv, 2a15, 1isi, 2c8a, 1yc2, 1xge, 1r0c, 1w8y, 1dl7, 7cpp, 1l4l, 1yum, 2g6f, 1qxo, 1ya6, 1i3a,
1sj1, 1uw6, 1u4o, 1o7s, 2ch5, 1w1a, 2goo, 1usr, 1o7p, 1jsz, 1gra, 1e7y, 1sqn, 1tcv, 1u3d, 2np5, 2g9k, 2gab, 1g38, 1u2o,
1qy5, 1g86, 1elv, 1ynl, 1c3o, 1vrt, 2g5u, 1y5w, 2efx, 1h2r, 1e4i, 1yki, 2dkd, 2a2c, 1bcj, 1ktc, 1d0h, 1wmz, 1fnz, 2a2d,
2ays, 2fyd, 1ax0, 1q3a, 1dqo, 1e6z, 2chn, 1hp5, 2f98, 2cuk, 2dsy, 1zhh, 1m5j, 1r6l, 1p7j, 1oks, 2ich, 2ite, 1h9x, 1cx9,
2ate, 1doh, 1zwp, 1jhq, 1icr, 2f7f, 1a6w, 1gzf, 2gjn, 6upj, 1lrh, 1zj1, 1mmt, 1eol, 1gs5, 2gu5, 1xpy, 1xcj, 1p1r, 1nus,
1hyb, 1ta8, 2gvg, 2hct, 1isj, 1d1c, 3pck, 1l9p, 2fbb, 1li4, 1lhv, 1dog, 1i75, 2j77, 1n5v, 1qqs, 1kic, 2fqw, 1a9s, 1pr0,
2bsx, 35c8, 2j75, 1ngp, 1a6v, 1sjd, 1kgq, 1asc, 1ls6, 2i10, 43ca, 1vah, 1z44, 2d20, 1yek, 1u68, 1o7g, 1zfk, 8nse, 2a5x,
1t9b, 1xgi, 1nis, 1ww3, 1qpq, 2b7n, 1i83, 1e6q, 1v08, 1noj, 2j7b, 1jdx, 2afz, 1aej, 1e1x, 1o8a, 2g5r, 1oxe, 1ovk, 1nmz,
5cts, 1oaa, 2h12, 1sgj, 1z6k, 2byo, 1c83, 1n5t, 2cbj, 1riv, 2gyw, 2de3, 1byf, 1byf, 1g8k, 1znh, 1v2g, 2bui, 1xl8, 1z03,
1y5m, 2gfc, 2j58, 2evd, 2gsk, 1cwq, 2bab, 2dm5, 1fh0, 2bac, 2oq7, 1s32, 2od6, 2uv0, 2inz, 2bj4, 2gpu, 2arm, 1mrf,
1r1j, 1j78, 1gni, 2ag4, 2ev1, 1gt6, 1lfo, 1hms, 2ftb, 1yh1, 1w0x, 4erk, 2cmw, 2d1r, 1ajp, 1km6, 1ofe, 1d0y, 1c86, 1q7a,
1icq, 1b7a, 1c3o, 3jdw, 1hqg, 1x7d, 1jqx, 2fqi, 1opr, 1j79, 1tv5, 1ep2, 1njj, 1fhv, 1sjb, 1c88, 2f1c, 2por, 2yhx, 2an4,
1lo2, 1xl0, 1r4u, 1ytm, 1o4n, 188l, 1hu0, 2ogf, 2dua, 1gz3, 1a49, 1pt8, 1nvm, 1pym, 1pzf, 1ayl, 1o4t, 1t2d, 1ldg, 9ldb,
1a5z, 1h17, 2d4q, 1ueh, 1ikt, 1t24, 1x8l, 1w3j, 2izu, 2gms, 2h7j, 2g2h, 1uog, 2ai2, 2ai1, 2hox, 2bxp, 1zg7, 1t9c, 1t9c,
1t9d, 2roy, 1u21, 2ai3, 2cgf, 1jhr, 1va6, 1zzm, 2aoa, 1oxn, 1y4l, 2iz1, 2aaw, 2hr7, 2ajs, 1wma, 2i02, 1xk9, 2bvc, 2j3u,
2f08, 8cho, 2b2n, 2gyu, 2ha2, 1w99, 2fxa, 2hzc, 1o57, 2fgb, 2inc, 2g81, 1g8i, 1n5q, 1x7n, 1tzc, 1g98, 2nr9, 1u0y, 2dza,
1iuu, 2ine, 1k5q, 1n2e, 2ofp, 2al1, 1l8p, 1g7v, 1jcx, 1ml4, 1fk3, 1uvb, 1izo, 2gc0, 1koj, 1oth, 2f9w, 1sq5, 1tjy, 1xuc,
1xur, 6cha, 1r9q, 1sw1, 1utp, 1e4h, 1fiw, 2a2q, 2bdg, 1okc, 1p6e, 1ym0, 1q3e, 1be4, 2gwh, 1y4z, 1mvn, 1p4a, 1g3m,
1jhv, 1xqx, 1tug, 1drt, 1gvg, 1mc1, 1elu, 3daa, 1c3v, 1l6g, 2bmk, 1kn4, 1x2a, 2gmu, 1l6f, 1h61, 2aax, 1zv9, 1iz8, 1ki3,
2fhj, 2g80, 1zb9, 2c12, 2cmz, 2gou, 1q74, 2hbo, 2fhj, 1z5p, 1y89, 1u3a, 1wma, 2oa5, 1jjb, 2bka, 1zej, 1znd, 1tnj, 1d6y,
1xio, 1ppj, 1yq3, 2c77, 2cfz, 2bi1, 2j0p, 2i5e, 1zgq, 1yq2, 1p7k, 2fbd, 2iz1, 1uk8, 1eyw, 1iwh, 2al1, 1zha, 1of8, 1vs1,
2fag, 1xuz, 2ox3, 1vbh, 2b7o, 2hml, 1pck, 1n46, 1mgo, 1you, 2hai, 1e7a, 2oat, 2br1, 2brb, 2cu9, 1ry0, 2caq, 1oe2,
1dx6, 1nzv, 1u0y, 1tc0, 2hkd, 1w2u, 2bly, 1o57, 1iwn, 1d1j, 1d1j, 1s8a, 2btm, 1tti, 1gg1, 1sw0, 1amk, 1pdz, 2bkv,
1lyx, 2fqg, 1rf2, 2i10, 2fxa, 1lrl, 2ccr, 1pjs, 2gq2, 2dry, 2g81, 1yei, 2isw, 1tph, 1ik4, 1ttj, 2cet, 2cer, 1tuk, 1ds1, 1icj,
1iye, 1x28, 1maq, 2irv, 1aj0, 1dm6, 2bve, 1d7l, 1q4s, 1fw9, 3pcc, 1o7t, 1g0n, 1tf9, 1nlu, 1lih, 1qpr, 1ehi, 1ocu, 1bzj,
2j7u, 2c13, 2bie, 2cek, 1uda, 1ozh, 2j0r, 1ur4, 2c41, 1uo5, 1xm4, 1phe, 1s1f, 1e9x, 2d0t, 1ecg, 2g09, 1v5d, 1cml, 1hfa,
2mas, 2nva, 1ajs, 1dfo, 1szr, 1lw5, 2jaf, 1o6u, 1o8v, 2ifb, 1mgp, 1b56, 1pz4, 1mzm, 2iu8, 1m66, 1q20, 1e5f, 1szs, 1l5v,
1yaa, 1em6, 1tzk, 2aq6, 1m32, 1wrv, 2dgm, 2ieg, 1pmo, 1beu, 1ft7, 1u19, 2okj, 1pye, 2gv2, 1fup, 1xzc, 1yp2, 1klk,

1x29, 1xql, 2o7e, 1kgt, 1eye, 2bmb, 1jhq, 1szs, 1aia, 1zc9, 1mdo, 2f8j, 2cjg, 1xql, 9aat, 1a0g, 2hoz, 1sh7, 1yec, 1yef, 1kn2, 1yej, 1gm7, 1zed, 1d0z, 1h1t, 1vlh, 1od6, 2a3c, 2c70, 1bpi, 3sil, 1yn9, 2fgv, 1pc3, 1lqk, 1okh, 1ex2, 2bwl, 2fhd, 1m32, 1sww, 1o72, 1h8p, 2ckq, 3lkf, 1woz, 1wkg, 1al4, 1al4, 1al4, 1al4, 1al4, 1al4, 1al4, 1al4, 1jse, 2au6, 1g29, 2oa6, 1uvk, 1g67, 2d4q, 1c9k, 1n22, 2bqy, 1vha, 2ivv, 1qcf, 1qpe, 1c9c, 1wc7, 1qpr, 1akb, 1arg, 1gey, 1akc, 1wkh, 1ei6, 1nki, 1m7y, 2hjp, 1kc7, 2ay4, 1vqp, 2bcu, 1qbv, 1p0b, 2bal, 2gfs, 2d0v, 1kb0, 1otw, 1yiq, 1cq1, 1f0s, 1tnk, 1ptr, 1fpu, 1boz, 1com, 1ozq, 1a4m, 2j3m, 1bcu, 1i14, 1kgz, 1opr, 1fsg, 1zvw, 1zyk, 1l1r, 7kme, 1pg3, 1owy, 1hqp, 1gt1, 1qy1, 1ykj, 1duv, 1br6, 1tx0, 1wm1, 1hqw, 1f8q, 1i2a, 1ykz, 2ae2, 1stb, 1jyq, 2cjz, 1pty, 1d1v, 2agv, 1uyh, 1uyg, 1uym, 1uy7, 1uy8, 1uy9, 1uyc, 1uyd, 1uye, 1uip, 1msv, 1a99, 1f3t, 2o06, 1uyk, 1uyi, 1w3y, 1b9i, 1td2, 1ho4, 1szr, 187l, 1py5, 1cq7, 1cq8, 2eur, 1m18, 1l9d, 2cwh, 1w61, 1kya, 1r16, 1ti4, 1qpb, 1ja9, 1g0o, 1w3o, 1aq2, 2ez8, 2c42, 1kbi, 2fm3, 1say, 2fn1, 1pj3, 2g50, 1ml7, 1n4g, 2bb7, 2jc6, 2c9z, 1gp6, 1h1i, 1jqe, 1p1o, 3cbs, 2cbs, 1z5l, 1xow, 1tgy, 2h4l, 1a9t, 2adu, 2be2, 1snn, 1zha, 1uj5, 1w6j, 1b9v, 1pw7, 1mex, 1b9t, 2e0s, 2fuq, 1xiv, 2d81, 1riu, 1t6z, 1kyv, 1nb9, 1bu5, 2ccb, 1l5r, 1l4e, 1x8t, 1e2n, 1fk7, 1evr, 2fxs, 2hkj, 1u0z, 2fyp, 1zwh, 1qy8, 1nqu, 2b99, 1cbq, 3lbd, 1tyr, 1gx9, 1fem, 1fm6, 1n4h, 2fr3, 1wor, 2iz1, 2bes, 1e0p, 1opb, 1ikg, 1iki, 1h66, 1jvi, 1rks, 1yoe, 4rhn, 1d0v, 1ogd, 2fn8, 2ioy, 1y7p, 1rqj, 2o1o, 1yv5, 1yhl, 1nqx, 2c67, 1jho, 1mdl, 2hwh, 1zmt, 1h46, 1de6, 1x8d, 2i56, 1de5, 1phw, 1r15, 1xmu, 1cbk, 1q9m, 1ne4, 1oge, 1qvj, 1px0, 1lij, 1g9v, 1unl, 1nfw, 1z9g, 1lox, 1kt7, 1crb, 1fmj, 1t3p, 1rcx, 1me8, 1zz8, 1b3d, 1wvb, 2iyz, 1mi4, 2ggd, 1rf6, 2o0l, 2cxq, 2gc1, 2c4t, 1z3c, 1gqr, 1xcl, 2c7p, 2aot, 1g55, 1nw7, 1or8, 2bzg, 2igq, 2hut, 2i62, 1im8, 2avn, 1xtp, 1y7i, 1fo4, 2fn1, 1jgs, 1g60, 2f8l, 2g72, 1wg8, 1jg4, 1nv8, 1mjq, 1vpt, 1n6a, 1qzz, 1aj0, 13gs, 3erk, 1bmk, 2ai7, 2ai8, 2aie, 1if9, 1x84, 1ah0, 1if7, 1if8, 1x8r, 2j1g, 2aj8, 1ynk, 2ha2, 1epv, 2daa, 1i2l, 1mdz, 1vfs, 1of1, 1hb3, 1zge, 1h0j, 1w9d, 1d5x, 2b6c, 2gkp, 1l7p, 1pbo, 1dzj, 4nos, 2bs3, 2c3p, 1siw, 1jnr, 2h88, 1h7x, 1oao, 2h9a, 1o94, 1nw6, 2h2j, 1qm5, 1qfm, 2iw9, 1za4, 1rsn, 4gsp, 1um4, 1ck6, 1zz1, 1ms7, 1lcw, 1e5q, 1hbm, 1rvz, 2qwb, 1e8u, 1qfo, 1vps, 1w0p, 2bf6, 2bat, 1mx1, 2c25, 1fuq, 2fdi, 2ayd, 1h6c, 2bwn, 1kor, 1f8i, 1cze, 3bif, 2og7, 1hg0, 2f6v, 2f7r, 1yz3, 2sli, 1hnn, 2iys, 2gpt, 2aay, 2d5c, 1zui, 1q36, 1v3c, 2ber, 1nxy, 1pi5, 2ffy, 1uka, 1x9d, 1sjc, 2g65, 1oo6, 1o9v, 2bwm, 2d41, 1s3f, 2ab2, 1zo8, 1dy4, 1ldo, 1lcv, 2g11, 2avo, 1enf, 2b6c, 1eyl, 1vqs, 1e7l, 1vb0, 2fi4, 2nac, 1va5, 1j2z, 1s5z, 3xim, 1d8c, 1gm9, 1ne6, 2ow9, 1ajq, 1fl3, 1d7x, 1pot, 2hmp, 1bo4, 2o07, 1pvc, 1mg9, 1se6, 1rfi, 2c12, 2b8j, 1rf4, 1vq8, 1m1b, 1kme, 1z5n, 1wle, 2ixi, 2ot9, 2c1l, 1um5, 1zjy, 1fyf, 1bto, 1q19, 1tqt, 1tqv, 1tqs, 1ivb, 1ivd, 1ivc, 1ive, 1inf, 1ing, 1inh, 1okn, 1l2s, 1fk4, 1lif, 1xxs, 1uvc, 1k4w, 1u0w, 1sg0, 1ya3, 1mrq, 2aa6, 2aba, 1x2b, 1u49, 1v8x, 1fee, 1muu, 1tn6, 1m98, 2gdu, 2hds, 1mw1, 1ynq, 1yni, 1k55, 1kf2, 1mz0, 1hlq, 1fn0, 1jo8, 1xvo, 1ynh, 1hww, 1wb4, 1sd3, 1wb5, 2bgd, 1w19, 2bge, 1w19, 1tha, 1zdy, 1z4k, 1cy8, 2rox, 1hk4, 1w19, 1w19, 1zuc, 1oum, 1q4n, 2iwv, 1s9a, 1tal, 2dm6, 1cb7, 1kwn, 1alu, 1s20, 1ei6, 1fs5, 1mze, 2dd4, 1ygy, 1n97, 2bjf, 1otj, 2bz1, 1vm1, 2hvk, 1f6t, 1e5a, 2bvl, 1dd4, 1p5e, 1j91, 1ruv, 2o78, 1iex, 1nhw, 1arc, 1uh5, 1qsg, 1nhg, 2b35, 8cpp, 1plf, 1j51, 1qll, 1lt5, 1a78, 1muq, 1y6q, 2ezt, 1x80, 1w88, 2o1s, 2iea, 1umc, 2fwn, 1iqu, 2j0f, 1kep, 2h0t, 1eyw, 2h88, 1q9i, 2a3a, 1q13, 2am9, 1vpo, 1jtv, 2bu5, 2bu6, 2bu7, 1n0q, 1g6u, 1w70, 1zzr, 1kb0, 1tiw, 1k2u, 1ctr, 2f2k, 2ajx, 2be5, 1xl1, 2cbo, 1j8u, 1df1, 2hr0, 2aid, 1p6x, 1w2g, 1p7c, 2b8t, 1h5r, 1kvl, 1a3u, 1cy1, 1jtk, 1qf1, 1qf2, 1aer, 1mmk, 1eoj, 1zdp, 1c5c, 1y59, 1y5a, 1y5b, 1y5u, 2aq7, 2aqb, 1k2y, 2gpt, 1oht, 2dbp, 2b13, 2cx1, 1gu1, 2fx5, 2dw6, 2d6y, 1fj4, 1lw4, 1e2l, 1jbw, 1mpw, 2bdm, 1tto, 1rf2, 1pbt, 1m6z, 1y7t, 2bme, 1xcd, 1tdh, 1jl0, 2bxw, 2o4g, 1n5k, 1e2d, 2ido, 2fto, 2ccg, 1h5s, 1cy2, 1z4l, 1oi6, 1ac4, 1q5k, 1ac8, 1n5q, 1xhl, 1vys, 1gvr, 2d3s, 1ae4, 1dg5, 1dyr, 1f4f, 2as1, 1hbo, 1tnl, 2ij7, 1bzc, 1f3d, 1lc7, 1y9d, 2c42, 1upa, 1trk, 1ni4, 1ozf, 1qgd, 1v5f, 2g1i, 1f4e, 1g4s, 1ia1, 1ia2, 1ia3, 1ia4, 1s3v, 1s3y, 1l5j, 1aco, 5acn, 2gsm, 1cwq, 2by0, 2e50, 1v6a, 2cy6, 1eu8, 2fpd, 2dxy, 2b1q, 2ebf, 1j1m, 2by0, 1hx1, 1ukk, 2fkk, 1zjb, 1d0i, 2cl0, 1qns, 1mvy, 1l5v, 1oiz, 1zgf, 2fp2, 1ecm, 1vql, 1t64, 1c3r, 2iuq, 1j4i, 1b0d, 1wuw, 1m51, 1hyv, 1xap, 1n4q, 1gz4, 1hyz, 1cr1, 1iim, 1njy, 1xjm, 1qtm, 1s0o, 1t3n, 1n5j, 2ggq, 1xbt, 2jg0, 2ivd, 1gzq, 1so4, 1txi, 1nlu, 1q11, 1h3f, 1n5l, 2ixi, 1lvw, 1kew, 1r66, 1epz, 1rrv, 1ket, 2ocu, 2dpz, 1tyl, 1w1y, 2fr5, 1mlz, 1dm8, 1rp7, 2fn3, 1ekk, 1jv4, 1g67, 1zzl, 2d2h, 1upj, 1rth, 1yf6, 2gg0, 2gg2, 2gg3, 2gg7, 2gg8, 2ggb, 2gg5, 2aix, 2ogz, 1eos, 1vd3, 1o0m, 1z4j, 1uca, 2bts, 4rsk, 1ucc, 1jsv, 2ffc, 1dbt, 1g8o, 1xtt, 2j4k, 2gui, 1hxp, 1ucd, 2bln, 1fgx, 2btr, 1w4o, 2i5x, 2f34, 2f35, 1s1t, 2d8l, 2ahg, 1r81, 1gy8, 1zi5, 2iyf, 2bgu, 2c9z, 2iya, 2gn8, 1nai, 1y6f, 1gz5, 1o29, 1sqo, 1sqt, 5upj, 1y3y, 1w4p, 2rnf, 1w4q, 1j3j, 1tsd, 1o25, 1njd, 1f7k, 1snf, 1seh, 2bsy, 2bl2, 2f6w, 1h0v, 2f70, 1e02, 1gt4, 1wrr, 1y9l, 1cwq, 2uvi, 1ueh, 1yba, 1vmg, 2g8l, 1vpg, 2f4p, 2ouw, 2i8d, 2ig6, 2hbw, 1vrm, 1vk8, 5std, 1km0, 2guu, 1lsh, 2d4n, 1zoy, 2i5n, 2prc, 1oe5, 1tgy, 1bd4, 1emj, 1brw, 1ucd, 1udh, 1ksk, 1ui0, 1gth, 1l5s, 1zab, 1upf, 1rxc, 1h7x, 1lnx, 2fr6, 1loj, 1uwk, 2fvm, 2icx, 2bnf, 2im2, 2ikf, 1r8c, 1twf, 2b56, 1ruv, 1jh7, 1e7q, 1uzi, 2aco, 2awh, 1rkg, 2jf4, 1mz9, 1j78, 1sbr, 1ig3, 1ig0, 2hh9, 1ohw, 2gfk, 1mmw, 1tuv, 2qr2, 2ahc, 2g9z, 1e59, 1nu3, 2cjp, 2hwi, 1rql, 2i1r, 1wb6, 1w06, 2fyv, 1s1s, 1wbu, 1v0r, 1j3k, 1dg7, 1o69, 2h5a, 2hgd, 1a96, 1rsc, 1oao, 1rco, 1j01, 2bmz, 5xin, 2brp, 2cu0, 1lol, 1qk5, 1mei, 2czf, 1pkx, 1xnk, 1xii, 2itm, 2fwq, 1pr6, 1l4l, 1kya, 1xsk, 1ec9, 1s5n, 1njr, 1w3y, 1px8, 1qh7, 1gqj, 2cdc, 1h12, 1goq, 2e2q, 2dua, 2hox, 1v6u, 1b3z, 2cn3, 1dwh, 3xis, 2fgl, 1wu5, 2bfg, 2ew5, 2ew6, 2gnj, 1mbx, 1t9d, 1t9b, 1vdv, 1gcz, 2bkl, 1lx6, 2flh, 1zfq, 2ewb, 1w1q, 2exm, 1gs4, 2htq, 1v3e, 2her, 2f9k, 2h2i, 1xc1, 2duz, 1mar, 1gwq, 2cfi

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form in einem Verfahren zur Erlangung eines akademischen Grades vorgelegt.

Saarbrücken, den

(Christoph Hartmann)