1. Introduction

Since Emil Fischer at the end of the 19th century recognized the "lock and key" principle for the interaction of small organic compounds and large biomolecules the research for new remedies has been focused on the non-covalent interactions between organic *keys* and protein *locks*. Until X-ray spectroscopy provided a first insight into the 3D structure of DNA [136], proteins [104] and protein/ligand complexes in the second half of the last century, new agents could only be detected by random trial, chemical intuition and the stepwise modification of already known active compounds. With the 3D pictures of the locks and keys it was, for the first time, possible to study the interactions between the ligands and their receptors at the molecular level. From now on the chemical mechanism of a certain pharmacological process could be investigated and better methods to find more efficient or new agents could be implemented.

Fischer's model was later refined by Koshland [75;76] who introduced the concept of "induced fit", which takes the flexibility of the ligand and the receptor into account. Thus, substrates and proteins initially do not fit into each other but are transformed into matching counterparts by conformational changes during the complexation process. Only after the substrate is bound, the partners have complementary shapes and form a lock and key pair. This model describes a process of dynamic recognition which enhances the selectivity of the protein. Nuclear magnetic resonance spectroscopy (NMR) can help to elucidate the flexibility of protein structures because it is able to reveal the atomic structure of a protein in solution [139]. Measurement of the nuclear Overhauser effect (NOE) allows the relative localization of atoms to each other in the three dimensional atomic structure. This technique is incorporated in nuclear Overhauser enhancement spectroscopy (NOESY), which - together with distance geometry - produces a set of possible 3D structures of the protein that can be interpreted as alternative conformations of the flexible protein [58;98].

Together with the evolution of protein theory the development of quantum chemistry [9] provided a basis for the theoretical investigation of the electronic structure of small molecules. With the advent of new computational chemistry methodologies the calculation and prediction of molecular properties of physically unavailable compounds became possible. But limited computer power and the enormous amount of computations that is necessary to perform *ab initio* quantum mechanical calculations prevented for decades the widespread use of theoretical methods in bioorganic chemistry. However, because of the dramatic increase in hardware performance in combination with less-demanding computational methods like semi-empirical quantum mechanics [40] or force field-based molecular mechanics [22], molecular modeling has become an integral part of the drug discovery process.

Since the end of the 20th century molecular modeling [79] has been defined as the collection of all theoretical methods that facilitate the prediction of molecular properties and activities by means of 3D atomic models. Superposition of 3D structures, alignment of molecular fields, docking of ligands into their receptors, *de novo* design and 3D-QSAR are typical tasks in a molecular modeling process. Together with modern computer graphics, now available in every commodity desktop computer, these methods can provide a detailed and very intuitive insight into macromolecular systems. Furthermore, molecular modeling in combination with distributed computing seems to be one of the few feasible approaches to investigate the vast amount of data that is created by the activities of the genome project [94;130].

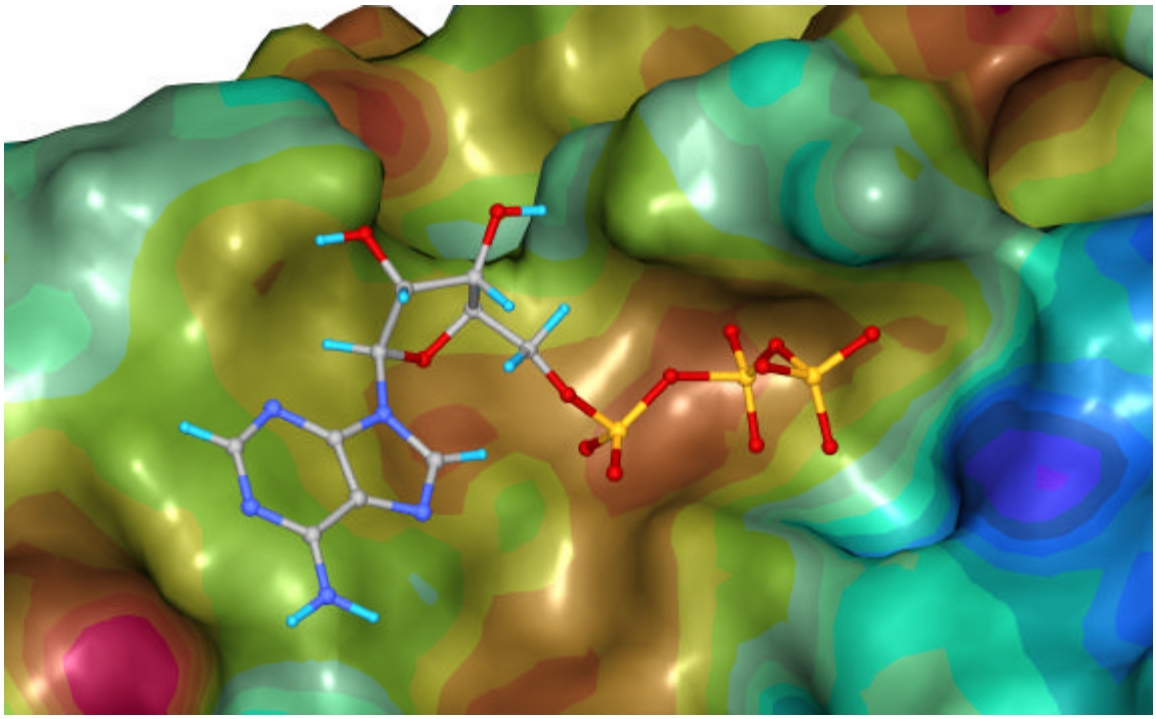

Figure 1-1: ATP binding site of the PDB entry 1B0U (ABC transporter protein).
The protein is represented by its surface, color-coded with the electrostatic potential (red=positive, blue=negative) and the ATP ligand is represented with balls & sticks in a CPK color scheme.

An important aspect in molecular modeling is the characterization of the non-covalent interactions between receptors and ligands which are mainly driven by hydrogen bonding, van der Waals forces and electrostatic fields. A sufficient complementary match between these features is necessary for two molecules to interact in a biochemical process: The teeth of the key must fit into the lock. Usually these interactions are analyzed by means of force field-based molecular mechanics methods [18] which can be very time consuming. However, it is evident that the physicochemical features at the surface of the molecules are more important than the properties of atoms buried deep inside a structure. Especially for large macromolecules the activities and properties are dominated by the features of their molecular surface [131], which can be illustrated very intuitively by color-coding the surface involved in binding with the relevant physicochemical properties (see Figure 1-1). Thus we can argue that the molecular surface augmented by physicochemical properties is a useful descriptor of the intermolecular non-covalent interactions. The affinity between two molecules can thus be understood by analyzing the complementarity of their surfaces involved in the binding process. By the same token, the ability of a compound to mimic the behavior of another one can be correlated to the similarity of their surfaces. But one should keep in mind, that molecular surfaces reflect only Fischer's lock and key model and cannot predict effects caused by induced fit. Therefore, investigations that involve surfaces are restricted to systems with a fixed geometry (e.g. already formed protein/ligand complexes).

1.1. Previous work

When Chothia and Janin showed that the complementarity of molecular surfaces plays a major role in the selectivity of protein/protein recognition [31], a proof of concept was established. Since then a large number of methods for the comparison of

protein or ligand surfaces have been developed which can be grouped into two main categories: the search for *complementarity* or *similarity.* Surface complementarity is one of the key aspects in molecular docking [34;102] and practically all docking methods include some sort of assessment of complementarity in their scoring functions. Surface similarity, on the other hand, is a valuable tool for the detection of common chemical features and is related to e.g. the active analog approach of Marshall et. al. [90]. Both complementarity and similarity-based methods can further be subdivided into algorithms that search either for *global* or *local* matches between the surfaces.

Some of the earlier methods were based on gnomonic projection or spherical parameter surfaces [15;17;30]. The common principle behind these methods is the mapping of the molecular surface onto a highly symmetric geometric object, such as a sphere or a platonic body. The similarities between different surfaces can then be examined by comparing the geometric objects instead of correlating the irregular original surfaces. Another way to globally compare the shape of two molecules is the use of Fourier shape descriptors [81;113]. In this case the surfaces are approximated by a series of spherical harmonic functions and represented by the corresponding coefficients. Both methodologies, the gnomonic projection and the Fourier analysis are inferior to other methods if the molecules are markedly non-spherical (i.e. have large and deep cavities). Correlation techniques, another kind of global methods, can deal with such shapes [72].

Especially in the field of molecular docking there is a need for local surface comparison, because the surface of a ligand, be it large or small, hardly ever fits into the complete site of a receptor molecule. For this purpose detection of local complementarity between two surfaces is essential. In a first attempt Connolly [34] searched for complementary groups of geometric features between two protein surfaces. He identified critical points – knobs and holes – on both surfaces and selected possible matches by a set of heuristics that checked the size and shape correlation between all knob-hole pairs. The initial implementation had the drawback that at least four positive matches were necessary to generate an alignment between the two surfaces. This issue was later solved by Wang [133] and Connolly [35]. The research group of Ruth Nussinov has refined the concept of critical points by the technique of geometric hashing [77] to enable a fast screening of the large set of possible matches in a protein/protein or protein/ligand docking run [51;86].

The innovations of Connolly and Nussinov et. al. represent important milestones in the development of surface comparison techniques. The idea to represent a complex surface by a small set of localized features lays the foundation of a new generation of molecular surface similarity or complementarity search algorithms. The concept is always the same: Reduced representations of the necessary surface features are compared by some heuristics and the matches are assembled to alignments by computer vision and graph-theoretical techniques, such as geometric hashing or maximum common subgraph isomorphism. Cosgrove et. al. introduced a shape based method that separates the surfaces into patches of approximately constant curvature and retrieves the surface similarities by clique detection [37]. Goldman and Wipke use quadratic shape descriptors (QSD) to represent the surfaces which are compared by their parameters. The matches are thereafter assembled by expansion of single QSD alignments [56]. Their method has also been adapted for docking [55]. Another remarkable approach is the surface segmentation of Heiden and Brickmann [60] where a molecular surface is divided into segments of similar chemical or geometrical character by means of fuzzy

logic [142]. Exner et. al. are using this principle for the identification of surface patterns [48] and docking purposes [49].

Besides the publications mentioned above many others have investigated the possibilities of molecular surface comparison (e.g. [10;96;105]). Good reviews on the topic have been published by Masek [92] and Via et. al. [131].

1.2. Concept

As described above, the investigation of molecular interactions by means of their surfaces can provide an important contribution to the understanding and prediction of chemical and biological activities. Surface similarity can help to identify compounds that have the same properties while surface complementarity can be used as a powerful tool for the prediction of protein/protein and protein/ligand interactions. In previous studies it has been shown that both tasks are strongly related to each other and a large number of methods provide both possibilities either explicitly or implicitly. It is also evident from the literature that local similarity is usually more important than global resemblance.

In the pharmaceutical industry one of the most important tasks is the fast screening of large compound libraries against biological targets to find possible lead candidates. In addition to the established experimental techniques, such as high throughput screening [7;47;69;95], molecular modeling becomes more and more important. Several papers have been published recently that investigate the possibility of high throughput docking in combination with protein structure prediction as a computational alternative to the expensive experimental screening techniques [5;42;43;50;129]. In this context molecular surface comparison could serve as an alternative or refinement of the pharmacophore screening of compound databases or the docking of small ligands into protein sites.

In the present doctoral project the primary aim was the development and implementation of an algorithm for the detection of local surface similarities based on shape and surface-mapped molecular properties. The approach, presented in this thesis, is based on graph theory and a computer vision technique called Harmonic Shape Image Matching [145] augmented by a sequence of filters to identify groups of corresponding points on two different molecular surfaces. Rigid-body alignment of the chemically similar surface regions can then be used to generate hypotheses about the common binding modes of a set of molecules. To deal with large datasets and result tables a scoring mechanism was implemented to enable the ranking of different molecules against a template.