5. Conclusion and Outlook

This thesis demonstrates that the comparison of molecular surfaces of small molecules and of protein active sites can be performed by a stepwise filtering algorithm. The relative alignments of several inhibitors in the active site of thermolysin could be reconstructed successfully with a quality comparable to other methods. Furthermore a scoring scheme could be established that allows the fast screening of a large result set and the comparative ranking of different surface comparison experiments. The same procedure is also applicable to the comparison of proteins if the search is restricted to specific regions of the surfaces. This allows the identification of differences in the binding modes of two SH2 domains to a phospho-tyrosine signaling peptide and to create a plausible alignment of two structurally unrelated but functionally related proteins.

5.1. The Advantages of SURFCOMP

All the experiments were possible because the implementation of the algorithm (SURFCOMP) did not only calculate a superposition based on surface similarity but allowed a much more detailed investigation of the matching regions on the different molecular surfaces. The ability to extract and display similar surface areas provides the means to check the reliability of the matches, to extend the similar surface patches and to correlate these to the molecular structure ultimately defining the similarities or differences between two molecules. Together with the local character of the search an overall picture can be build, which contains all possible combinations of local surface similarities between two molecular shapes. With this detailed information it is possible to perform different experiments such as searching for similarities and dissimilarities or aligning two molecules based on their similar surface patches. This detailed investigation of molecular surfaces, however, is slower than other methods like the quadratic shape descriptors (QSD) [56] or SPAt [37]. On the other hand the consensus scoring methodology supports a fast screening of the results which is useful for the examination of large sets of alternative surface alignments that can be produced by the comparison of proteins.

The filter based procedure of SURFCOMP also provides a flexible framework that can be adapted to a large variety of surface similarity problems. It is possible to arrange the tests that are performed by the fuzzy, harmonic map, distance and overlap filters in a different way. For example, more than one chemical property can be checked by the fuzzy similarity function or the harmonic maps can be used for the shape and the chemical properties. The system can also be extended very easily. If additional checks seem to be necessary or further modification of the input and output data should be applied, one can add new filters and processing steps to the framework.

In contrast to other surface comparison methods like the QSD [56], SPAt [37] or the surface segmentation of Exner et. al. [48] SURFCOMP does not rely on a specific representation of the local surface patches. In the present experiments only circular patches were used but all the different filters that are applied do not rely on that concept. It is possible to use different surface patches such as the segmented surfaces or patches that are based on functional groups. The same is true for the selection of the critical points. The program uses only “peaks” and “valleys” as centers of the surface patches but the selection procedure can be extended to choose also saddle points or extreme values of various physicochemical properties.

SURFCOMP is also applicable to the comparison of at least parts of protein surfaces. Although the comparison of parts of molecular surfaces can possibly be performed by various other programs this thesis shows for the first time that a detailed investigation of similar and dissimilar patches on a protein surface can lead to interesting results for drug discovery and function prediction.

5.2. Discussion

Although surface comparison can be very illustrative, the simplicity and beauty of the pictures can blind the observer. A molecular surface is a very simplified model of a chemical compound and therefore provides only a limited view of biochemical processes. The German language has the right words to illustrate that restriction: The corresponding adverb for *superficial*, "oberflächlich" has the same roots as "Oberfläche" which means *surface*. A surface is always only a reduced representation model of the corresponding object. An old building, for example, might have some nice balconies and a fresh painting which make it look beautiful and well preserved but you can only confirm that impression if you check the rooms inside, the electric installations or the plumbing. The same is true for molecular surfaces. A large negative patch might indicate the presence of a nucleophilic agent, but you can never be sure until you look behind the surface at the structure of the molecule.

The flexibility of molecules makes the situation even more complicated because the shape of a compound can vary dynamically when adopting different conformations as illustrated in section 4.1.5. Already minor changes in the 3D conformation are sufficient to change the surface considerably. The surface does not contain the information any more that would be necessary to track the rotations, bending and stretching that cause these effects. For that it is again necessary to look beneath the surface at the atoms and bonds which provide the right model for that purpose.

Nevertheless, superficiality has also some advantages that make surface-only comparison of objects extremely useful. In 1984, many of the scenes in the famous motion picture *Amadeus* by Milos Forman [52] were taken in Prague, although most of the story was located in Vienna. To give the audience an impression of Mozart's life, the director had to find a place that looked like the capital of Austria in the second half of the 18^{th} century. He could not film in Vienna itself because too much had changed in the last 200 years. But some parts of Prague still had the typical buildings and streets of the time and similar facades or surfaces were sufficient to reconstruct the sight of Mozart's neighborhood. Similarly, when designing pharmaceutical compounds, reproducing the shape and the physicochemical properties of the original ligand is often a very successful strategy.

Most of the surface comparison experiments in this thesis followed this *look-alike* principle. Good examples are the common surface patches that were identified during the comparisons of different inhibitors and substrates for the thermolysin and dihydrofolate reductase. Those that have been found between all molecules of a set are likely to contain the necessary shape and electrostatic features that are recognized by the receptor. It is possible that these features are due to different functional groups, like the carboxylic and phospho-groups in the thermolysin inhibitors or the different heterocyclic rings in the DHFR ligands. These differences in the underlying structure will not influence the result of the comparisons as long as they manifest themselves in similar shapes and physicochemical properties.

Obviously the surface comparison methodology presented in this thesis can be successful only if the chemically relevant properties are mapped onto the surface. For

instance in thermolysin a Zn- ion is a key element of the active site and all known inhibitors are blocking this ion via a chelate complex. If a molecule has a functional group that generates a negative ESP patch similar to a negative patch due to a carboxylic group, but does not form a chelate, the surface comparison based on ESP may identify the compound as similar but it might not be active at all. Nevertheless, if a good model of the function of the active site is available and the physicochemical details of the ligand-receptor interaction are known, then a surface comparison can be more successful than a simple structure similarity search. In the latter case it would be difficult to figure out all possible combination of functional groups that cause these effects in advance.

The SURFCOMP program could find similarities between the surfaces of different SH2 domains and between the protein SAP and a tyrosine phosphatase (PTP1B) that have similar activities. The conclusion of these experiments is that common physicochemical surface patterns seem to be necessary for different active sites to show the same biological function. Like in the comparison of surfaces of small molecules different functional groups or residues can give rise to similar patches. Sometimes these residues are different but closely related to each other and sometimes they are totally unrelated. This is important, because it underlines the necessity of structural or surface studies between proteins. The question that remains is whether common surface motifs are not only necessary but also sufficient for similar functions of different proteins? For enzymes, surface similarities in the active site are certainly not sufficient because the mechanism of the catalysis requires well-defined side chains that are usually extremely conserved across a given enzyme family. For receptors, where non-covalent interactions between receptor and ligand dominate the recognition process, the actual chemical nature of the functional groups in the binding site is less important. In these cases it is sufficient if the surface of the binding site shows the physicochemical surface pattern necessary for specific ligand binding.

The comparison of SAP and EAT-2 showed that it can be rewarding to look for dissimilarities between the surfaces of active sites with similar functions in order to find ways to selectively influence one target molecule over the other which is often a very important problem in rational drug design. With a sequence or structural alignment only the differences in the amino acid sequences or the atomic positions can be detected. Molecular surface comparison can make the influence of these variations on the interface between the receptor and the ligand visible. One can then focus on those dissimilarities in the sequences that are responsible for the significant differences detected between the binding site surfaces.

Another benefit of protein surface comparison is the alignment of similar surface patches that is automatically created by SURFCOMP and can be used to establish a superposition of the complete protein structure and surface. In the comparison of SAP and PTP1B it was shown that a meaningful alignment could be constructed based on active site surface similarities, whereas the sequences and the 3D structures of the two molecules could not be aligned properly. In that particular case the surface alignment was reasonable because it resulted in a similar relative orientation between the active sites and the corresponding ligands. In general such an alignment is complementary to sequence and structural alignments. It does not focus on atomic and residue coordinates but on the physicochemical features of the surface points and can thus highlight functionally important similarities.

In summary, the power of a surface comparison lies in the highlighting of molecular properties closely associated with intermolecular interactions. The most

important limitation of surface comparisons is their lack of predictive power if molecular flexibility or chemical reactions play an important role.

5.3. Outlook

Molecular surface comparison is a rather new topic and only a few applications in drug discovery or molecular modeling have been established so far. In the present doctoral project no time was left to discover and tune all the possibilities of the new methods although several extensions and improvements are conceivable.

In section 2.2 different molecular and atomic properties were discussed that can be mapped on the points of a molecular surface. In the experiments mainly the electrostatic potential was used, but surface comparisons are not restricted to the ESP nor to the properties mentioned before. Different problems usually require different surface properties and one should select them carefully to meet the current requirements and models. Furthermore, as mentioned above, it is possible to use them not only in the fuzzy filter but also in the harmonic image step. This would provide information about the physicochemical similarity not only at the critical points, which was sufficient enough for the present experiments, but in the entire neighborhood of the CP. When applying that modification, one should keep in mind that the fuzzy filter is much faster than the harmonic images.

In the literature many docking algorithms that use surface complementarity are known [34;49;51;55;103;134]. Hence SURFCOMP should be applicable to docking tasks as well. Unfortunately, some preliminary tests, where the inverted surface of the DHFR receptor was compared with methotrexate, did not find any positive hits. The author believes that various steps in the framework, especially the harmonic shape image and distance filters, could not cope with the different sizes between the negative receptor and positive ligand surface. A possible solution to this problem could be to scale one of the surfaces so that the gap disappears or is reduced. In that case the distances would become comparable and a proper docking of the ligand into the receptor could be achieved by SURFCOMP. Scaling could be achieved by simple rigid body transformation or by the use of larger van der Waals radii in the generation of the ligand surface. It is conceivable that with these and similar modifications SURFCOMP could be adapted to function as a scoring component in a docking program.

One of the most interesting applications for drug discovery would be the comparison of compound databases against a set of known ligands to find possible antagonists and inhibitors. For that purpose it is necessary to perform and evaluate a large number of surface comparisons and to consider conformational flexibility. It is assumed that the method is fast enough to cope with a large amount of similarity searches. Such high-throughput applications can be “parallelized” easily on a large Linux cluster or on a distributed metaprocessor system such as United Devices [4] without any modifications of the system because each comparison is a single independent run of the SURFCOMP program. The evaluation of the results can be performed very rapidly by the consensus scoring method as described in section 3.10 which enables a fast screening and ranking of many compounds.

Incorporation of conformational flexibility is more difficult but not impossible. Although the method is only applicable to rigid 3D structures it is possible to combine it with a conformational analysis and to scan a set of low energy conformations of each molecule as expected from a complete 3D molecular similarity analysis. For that purpose one has to generate a database that contains molecular surfaces of several representative conformations of every entry in the original set of compounds and

compare this database against the template surface. This will increase the number of similarity searches linearly by the number of coordinate sets that are stored for each molecule. A simpler but less reliable alternative would be to generate a set of conformations for the template molecule and compare it against a set of rigid query compounds. In this case any positive hit must be checked against the natural conformation of the template to ensure that a low energy conformation of the hit structure matches the binding conformation.

Another interesting application would be the mapping of an unknown binding site by means of investigating known binders. If a set of compounds is known to be substrates or antagonists of a specific protein one can try to find common surface motifs on these molecules that may reveal pharmacophoric features which are necessary for the molecular recognition in that system. SURFCOMP provides the methodology to detect common patches between pairs of surfaces. By comparing one surface of a set with every other surface it is not difficult to select those patches that are similar between all molecules (see also Figure 4-4 on p. 46). These patches can then serve as negative images of the features that are present in the active site (e.g. a concave, electrostatic positive patch will most probably be matched by a convex, negative surface patch in the receptor). For such experiments conformational flexibility is essential because it will not be possible to determine the correct ligand conformations in the active site. It can be incorporated in the same way as described in the last paragraph, but in that particular case a comparison between all conformations of all molecules in the set will be necessary causing a quadratic increase of the pairwise comparisons.

According to Via et. al. [131], it should be possible to identify proteins with common functions by common surface motifs. Finding similar surface patches on structurally unrelated proteins was one of the motivations to start this project. Unfortunately, the task proved to be more difficult than initially expected. The main difficulty is the identification of the relevant sites on the protein surfaces because a complete comparison would be too time consuming and would produce too many results. Furthermore it must be clarified for each protein structure which crystal water should be considered as part of the structure and whether the sidechains of the amino acids should be relaxed or not. This process involves a lot of manual interaction and chemical intuition which is a rather time consuming process for the whole set of solved protein structures. But if all these problems can be solved SURFCOMP will be able to identify common motifs on all or some surfaces of the known protein structures which may reveal functional connections between unrelated protein families.