Die approbierte Originalversion dieser Dissertation ist an der Hauptbibliothek der Technischen Universität Wien aufgestellt (http://www.ub.tuwien.ac.at).

The approved original version of this thesis is available at the main library of the Vienna University of Technology (http://www.ub.tuwien.ac.at/englweb/).

Dissertation

Molecular Surface Comparison. A Versatile Drug Discovery Tool

ausgeführt zum Zwecke der Erlangung des akademischen Grades
eines Doktors der technischen Wissenschaften unter der Leitung von

Prof. Dr. Hans Lohninger
164
Institut für Chemische Technologien und Analytik

eingereicht an der Technischen Universität Wien
Technisch-Naturwissenschaftliche Fakultät

von

Dipl.-Ing. Christian Hofbauer
Matr. Nr. 94 09 824

Gentzgasse 15/II/28
A-1180 WIEN

Wien, am 31. August 2004

Zusammenfassung

Im modernen Entwicklungsprozess für neue Medikamente spielt der Einsatz von theoretischen und graphischen Methoden eine immer bedeutendere Rolle. Vor allem bei der Suche nach potentiellen Wirkstoffen für spezifische Proteine ist eine umfassende Kenntnis der Struktur des Proteinrezeptors von entscheidender Bedeutung. Neben der dreidimensionalen atomaren Zusammensetzung, wie sie von Röntgenstrukturanalyse und NMR Spektroskopie ermittelt werden kann, sind Oberflächendarstellungen besonders dazu geeignet, die Form und räumliche Ausdehnung eines Moleküls wiederzugeben. Zusätzlich lässt sich die Verteilung verschiedener chemischer und physikalischer Eigenschaften auf Moleküloberflächen sehr intuitiv darstellen und für optische Vergleiche verschiedener Strukturen einsetzen. Um die Analyse großer Substanzdatenbanken zu erleichtern, ist es allerdings notwendig die Suche nach ähnlichen Motiven auf Moleküloberflächen zu automatisieren. In der vorliegenden Arbeit wird das Computerprogramm SURFCOMP vorgestellt, das in der Lage ist, die Oberflächen verschiedener chemischer Verbindungen miteinander zu vergleichen und die gemeinsamen oder auch unterschiedlichen Eigenschaften zu ermitteln.

Da die Anzahl der Elemente, aus denen sich die Oberfläche eines Moleküls zusammensetzt, um ein Vielfaches höher ist als die Anzahl seiner Atome und Bindungen, ist es notwendig eine Darstellung zu wählen, die mit wenigen Elementen die charakteristischen Eigenschaften der Moleküloberfläche ausreichend beschreibt. In SURFCOMP werden dazu alle kritischen Punkte auf den beiden Oberflächen verwendet. Diese liegen entweder an der Spitze einer konvexen Region (Hügel) oder am Boden einer konkaven Region (Tal). Mit diesen Punkten wird nun ein Assoziationsgraph gebildet, der alle potentiell ähnlichen kritischen Punktpaare beider Oberflächen enthält. Dieser wird mit Hilfe von mehreren Filtern, die jene Paare eliminieren, die entweder aus chemischer oder geometrischer Sicht nicht zusammenpassen, soweit vereinfacht, dass gemeinsame Motive mit Hilfe einer Cliquensuche erkannt werden können. Die dabei detektierten Ähnlichkeiten sind überwiegend lokaler Natur. Um ein umfassendes Bild aller möglichen Gemeinsamkeiten auf beiden Oberflächen zu erhalten wird abschließend eine hierarchische Clusteranalyse aller gefundenen lokalen Ähnlichkeiten durchgeführt.

Mit dem vorliegenden Programm konnte zunächst die relative Orientierung von acht verschiedenen Inhibitoren im Rezeptor von Thermolysin erfolgreich reproduziert werden, was durch den Vergleich mit bereits publizierten Programmen belegt wurde. In weiterer Folge wurden die Auswirkungen von unterschiedlichen Algorithmen zur Generierung von Moleküloberflächen auf die Ergebnisse untersucht und die Flexibilität der hier vorgestellten Methode auf Konformationsänderungen der Moleküle getestet. Schlussendlich konnte SURFCOMP erfolgreich zum Vergleich von Proteinoberflächen eingesetzt werden. Bei einer Gegenüberstellung zweier SH2 Domänen (SAP und EAT-2), die beide an dasselbe Signalpeptid gebunden waren, konnte eine Reihe von oberflächlichen Unterschieden auf Differenzen in der Aminosäuresequenz zurückgeführt werden. Der Nachweis von ähnlichen Motiven auf den Oberflächen der reaktiven Zentren von SAP und der Phosphatase PTP1B konnte die in biologischen Experimenten entdeckte Aktivität von SAP zur Dephosphorylierung von Phosphotyrosin bestätigen.

Abstract

Analysis of the distributions of physicochemical properties mapped onto molecular surfaces can highlight important similarities or differences between compound classes, contributing to rational drug design efforts [131]. This thesis will present a method that uses a combination of graph theory, computer vision and computational chemistry to detect local surface similarities between small and medium sized molecules. The present approach is based on 3D structure search where maximal common subgraph isomorphism is used to detect local similarities between the pharmacophoric feature points of different molecules [91]. The extension of this principle to molecular surfaces is cumbersome, because treatment of the complete set of surface points instead of just a few feature points with NP-hard graph algorithms is not feasible. In order to perform a reliable and fast detection of local surface similarities it is necessary to reduce the complexity of the problem by a set of filters that implement various geometric and physico-chemical heuristics.

To achieve this, a simplified representation of the surfaces is generated first consisting only of a set of critical points (corresponding to "hills" and "valleys" on the surface), augmented by their surrounding surface patches. Among all possible point pairs those are selected first that show sufficient chemical similarity, judged by means of a fuzzy dissimilarity index [48] between physicochemical properties mapped onto the surface points. Then the curvature patterns around all remaining point pairs are compared by harmonic shape image matching [145] to discard points that are not embedded in a similar shape. Finally the distances and angles between combinations of similar pairs are checked to be within certain boundaries to form an association graph that is simple enough for the clique detection. The cliques represent the local surface similarities and an alignment between the two molecular surfaces can be calculated based on the corresponding points. Finally the alignments can be clustered to reveal a picture of the total surface similarity between the two molecules.

The method was tested with a dataset of eight thermolysin inhibitors and recovered the correct alignments of the compounds bound in the active sites. The results were in good agreement with another surface-based comparison carried out on the same dataset [37]. Furthermore SURFCOMP was successfully applied to the comparison of protein active sites by means of spherical site selection and a scoring scheme that allows a fast identification of similar surface regions. A similarity search between the binding area of two similar SH2 domains (SAP and EAT-2) revealed interesting differences between their molecular surfaces, which could be assigned to the corresponding structural differences. Finally the surface similarities between SAP and tyrosine phosphatase PTP1B, which have been detected by SURFCOMP, support the idea that biological functions are strongly related to surface features, since SAP and PTP1B do not show any significant structural similarity.

Acknowledgements

This project would have never been successful without the support of my direct supervisor Dr. András Aszódi, to whom I want to express special thanks. His door and mind were always open for the discussion of emerging problems and new ideas and he managed to control the progress of my work while providing me with the utmost possible freedom. Besides András I want to thank my academic supervisor Prof. Dr. Hans Lohninger for his support of the project and his contribution to the publication of the method.

I would like to thank all my present and former colleagues from the In Silico Sciences Group at the Novartis Institute for Biomedical Research in Vienna – including Benjamin Almeida from the IT department – who I enjoyed working with and who helped me in many fruitful discussions that fostered the success of my project. I also want to acknowledge the help of Dr. Tamas Schweighoffer who provided me with insight into the biological problems related to the SH2 and phosphatase experiments. In addition I want to thank Drs Adrienne James and Torsten Schindler for revising parts of the manuscripts of the thesis and of the submitted publications.

Finally, I want to thank my family for their unbroken support and love.

Table of Contents

1. Introduction 1
    - 1.1. Previous work 2
    - 1.2. Concept 4
2. Theory 5
    - 2.1. Molecular Surfaces 5
        - 2.1.1. Background 5
        - 2.1.2. Van der Waals Surface 5
        - 2.1.3. Isodensity Surface 6
        - 2.1.4. Solvent-Excluded and Solvent-Accessible Surfaces 8
        - 2.1.5. Representation of Molecular Surfaces 9
    - 2.2. Molecular Surface Properties 10
        - 2.2.1. Molecular Potentials 11
        - 2.2.2. Atomic Properties 12
        - 2.2.3. Surface Characteristics 12
    - 2.3. Feature Radius and Auto Correlation 14
    - 2.4. Fuzzy Logic 15
    - 2.5. Harmonic Images 17
        - 2.5.1. Concept 17
        - 2.5.2. Border Mapping 17
        - 2.5.3. Interior Mapping 18
        - 2.5.4. Generation and Comparison of Harmonic Shape Images 20
    - 2.6. Maximum Common Subgraph Isomorphism 23
3. Methods 25
    - 3.1. General Concept 26
    - 3.2. Definition of Critical Points 28
    - 3.3. The Association Graph 29
    - 3.4. Generation of Surface Patches 29
    - 3.5. Fuzzy Filter 31
    - 3.6. Harmonic Shape Image Filter 31
    - 3.7. Distance Filter 33
    - 3.8. Overlap Filter 34
    - 3.9. Clique Detection and Clustering 35
    - 3.10. Scoring and Ranking 35
    - 3.11. Treatment of Protein Surfaces 37
    - 3.12. Implementation Details 38
        - 3.12.1. Software 38

3.12.2. Hardware and Computation 38

4. Computations and Results 40
  - 4.1. Ligand Surfaces 40
    - 4.1.1. Preparation of the Input Structures and Experimental Design 40
    - 4.1.2. SURFCOMP Validation: Comparison of 8 Thermolysin Inhibitors 40
    - 4.1.3. Ranking of Surface Alignments 47
    - 4.1.4. Evaluation of Different Surface Types: Comparing DHFR ligands 55
    - 4.1.5. Testing different conformations 61
  - 4.2. Comparing Proteins: Surface Differences between SAP and EAT-2 67
    - 4.2.1. Surface Comparison 70
    - 4.2.2. Structural Investigations 72
  - 4.3. Elucidating the Phosphatase Activity of SAP 77

5. Conclusion and Outlook 82
  - 5.1. The Advantages of SURFCOMP 82
  - 5.2. Discussion 83
  - 5.3. Outlook 85

A The SURFCOMP Program Suite 87
  - A.1 Requirements 87
  - A.2 Availability 87

B Publications 88

C Abbreviations 90

D Indices 92
  - D.1 Figure Index 92
  - D.2 Table Index 93
  - D.3 Chart Index 93

E References 94

"Research is endlessly seductive, but writing is hard work"

Barbara Tuchman, *The Guns of August.*

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].

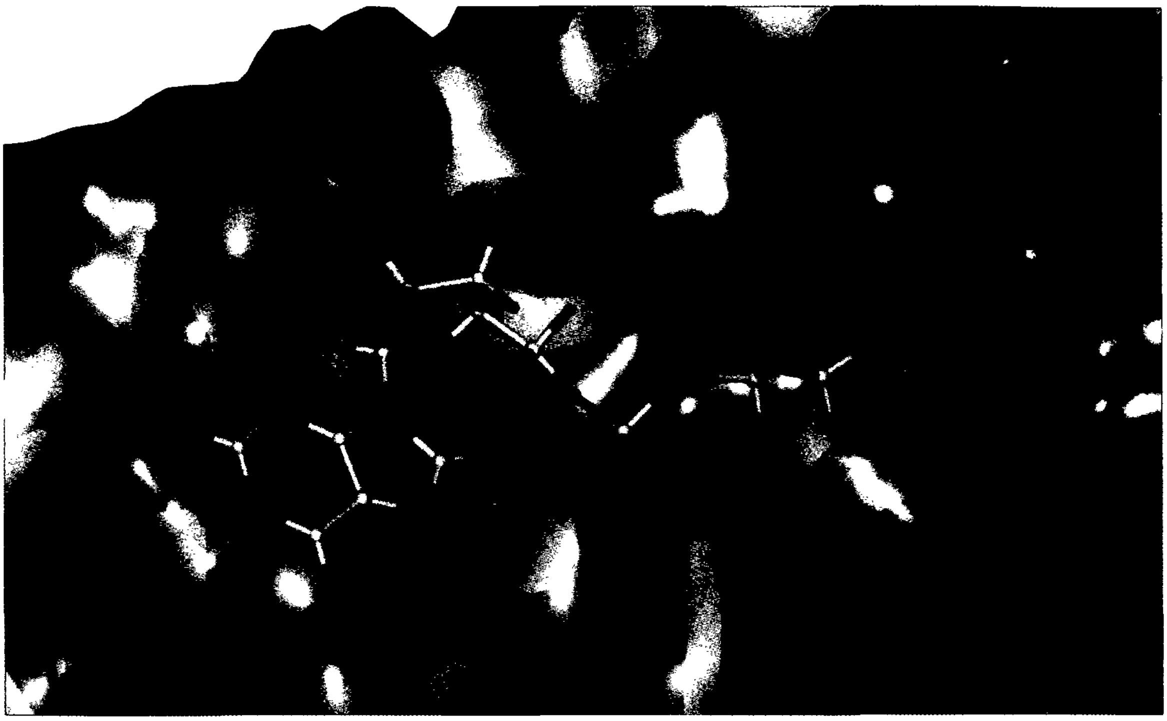


Figure 1-1: ATP binding site of the PDB entry 1B0U (ABC transporter protein).
The protein is represented by its surface, color-coded with the electrostatic potential (red=positive, blue=negative) and the ATP ligand is represented with balls & sticks in a CPK color scheme.

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.

2. Theory

2.1. Molecular Surfaces

2.1.1. Background

Since the days of Friedrich August Kekule graphical descriptions of molecular structures are widely used: Structural formulas just describe the topology of the molecule's atoms, which is important for the reactivity and can explain most reaction mechanisms. With the development of stereochemistry the combination of topology and 3D atomic coordinates gained importance. In a usual 3D structural formula of a molecule atoms are represented by dimensionless points and bonds are described by simple lines between these points. But molecules definitely have a spatial extension that can be described in various ways, and that extension is in many cases important for the outcome of a reaction or biochemical process.

Since molecules and especially atoms are very small objects, they fall into the realm of quantum mechanics where an absolute description of a molecule is not possible because of the uncertainty principle. Thus a border that defines what is inside and outside of a molecule is not as easily defined as e.g. for a rubber duck. Nevertheless since the work of Johannes Diderik van der Waals who investigated the influence of atomic and molecular volumes on the behavior of real gases [128], a large number of theories and definitions for the volumetric extension of atoms and molecules have been described. In this context the molecular surface can be defined as the boundary outside of which the molecule shows only weak non-covalent interactions with another molecule. The following subsections will describe the most important definitions for a molecular surface.

2.1.2. Van der Waals Surface

The deviation of real gases from the ideal behavior, as expressed in the van der Waals equation for real gases [127], is perhaps the first indication of a molecular and atomic volume:

$$(p-\frac{a}{V^2})\cdot(V-b)=R\cdot T \tag*{eq. 2-1}$$

where p is the pressure of the gas, a is a measure of the attraction between the particles, V is the volume of the gas per mol, b is the total volume of a mol particles, R is the ideal gas constant and T is the absolute temperature. A second evidence is X-ray crystallography of rare gas crystals. According to these results, each class of atoms (elements) can be modeled as a hard-sphere with a well defined radius, the so called van der Waals radius.

At short distances the repulsion between two atoms increases rapidly. This is due to the partial overlap of their electron clouds which causes a conflict with the Pauli principle. At medium distances fluctuations in the electron clouds are inducing dipoles in neighboring clouds which lead to a minimum in the potential energy. The van der Waals radius can be interpreted as the half of the distance between two atoms (of the same chemical element type) where the attractive mid-range forces are exactly balanced by the short-range repulsion. The radii can be determined experimentally from neighbor-neighbor interactions in crystals and from gas critical volumes [20]. In molecular mechanics calculations the van der Waals energy is usually described by the Lennard-Jones potential:

$$E(\mathbf{r}) = 4\varepsilon \cdot \left[\left(\frac{\sigma}{\mathbf{r}} \right)^{12} - \left(\frac{\sigma}{\mathbf{r}} \right)^{6} \right] \qquad \text{eq. 2-2}$$

where **r** is the interatomic distance and ε as well as σ are experimental fitting constants.

If each atom in a molecule is represented by its van der Waals sphere the space around a molecule can be divided into regions of mainly covalent and non-covalent interactions respectively. The interface between these two regions is the set union of all sphere surfaces that are not within any other atom's van der Waals sphere. This surface is called the van der Waals surface. It consists of a set of calotte faces around the atoms which are connected by circles that are located over the bonds.

The van der Waals surface is the simplest definition of a molecular surface and can be very useful when one investigates the effects of non-covalent interactions such as electrostatics or sterical clashes between two molecules in close contact. Its simple representation by a set of spheres provides the means for a fast decision if a point has to be considered inside or outside of a molecule. But more complex forms are also possible. Whitley, for example, developed a van der Waals surface graph, where vertices represent calottes, and edges between two vertices correspond to a circle connecting two calottes [138]. This graph can be used to study and describe molecular shape.

A disadvantage of this kind of surface is that it does not provide much more information than a 3D molecular structure. It just gives the single atoms in the molecule a volume. Other definition of molecular surfaces, as described in the sections below, provide additional information like the location of a specific electron density level or the volume that is excluded by a solvent molecule.

2.1.3. Isodensity Surface

From the quantum mechanical point of view a molecule is a set of bare nuclei surrounded by a fleet of electrons that are placed in specific molecular spin orbitals. Because of the uncertainty principle it is not possible to localize each single electron exactly, so an orbital is just a probability distribution over space that specifies where it is most likely to find an electron that is associated with it. The probability is expressed as the square of the function that mathematically describes the spin orbital. This square is normalized, so that the probability to find the electron in the complete space is equal to one:

$$p_i(\mathbf{r}) = \int |\chi_i|^2 \cdot d\mathbf{r} = 1 \qquad \text{eq. 2-3}$$

where $p_i(\mathbf{r})$ is the probability to find an electron of orbital i at the position **r** and χ_i is the spin orbital function. According to the Born interpretation of the wavefunction, the electron density distribution, $\rho(\mathbf{r})$, of the whole molecule can be interpreted as the probability to find an electron at any given point around the nuclei. This probability is the sum of squares of all spin orbitals that form the wave function of the molecule:

$$\rho(\mathbf{r}) = 2\sum_{i=1}^{N/2} |\chi_i(\mathbf{r})|^2 . \qquad \text{eq. 2-4}$$

The summation is over all N/2 doubly occupied spin orbitals and has to be counted twice because of the double occupancy. The spin orbitals and electronic wavefunction are usually calculated by *ab initio* or semi empirical calculations.

The electron density is highest near the nuclei and decreases with increasing distance. If we take a low threshold level of the density, the interface between regions that have more and less electron density than this threshold are separated by a smooth surface that encloses all atoms of the molecule. This is the most fundamental form of a molecular surface definition because it is directly based on quantum chemistry. Figure 2-1 shows the shape of the isodensity levels in 2D on a plane through the adenosine-triphosphate molecule.

The analogy between the van der Waals and the isodensity surface can be found in the definition of the van der Waals radius as the half-distance between two atoms where the repulsion and attraction of the van der Waals interactions is equal. The repulsion, as mentioned above, is caused by a violation of the Pauli principle due to overlapping electron clouds. Considering that a certain amount of electron density is necessary to make this effect significant, the isodensity surface can be seen as an extension to the van der Waals surface which describes that barrier for the molecule as a whole and not by means of a sum of hard-sphere atoms.

In addition to the complete electron density map, isodensity surfaces can also be generated for the probability distributions of single molecular orbitals or combinations of those orbitals. Of particularly interest for the reactivity of a molecule are the shapes of its HOMO and LUMO. However single orbitals are not representing the total shape of the molecule and their isodensity representations should not be considered as molecular surfaces.

A big disadvantage of the isodensity surface is its dependence to quantum mechanical calculations which are in general very time consuming and often restricted to small problems. The calculation of a large biomolecule is not feasible by most quantum chemical methods and electron density surfaces are thus available for small molecules only. Brickmann and coworkers have therefore implemented a fast calculation for electron density into their MOLCAD package [24]. The electron density of the molecule is approximated in the following way:

The electron densities are described by exponential functions placed on the center

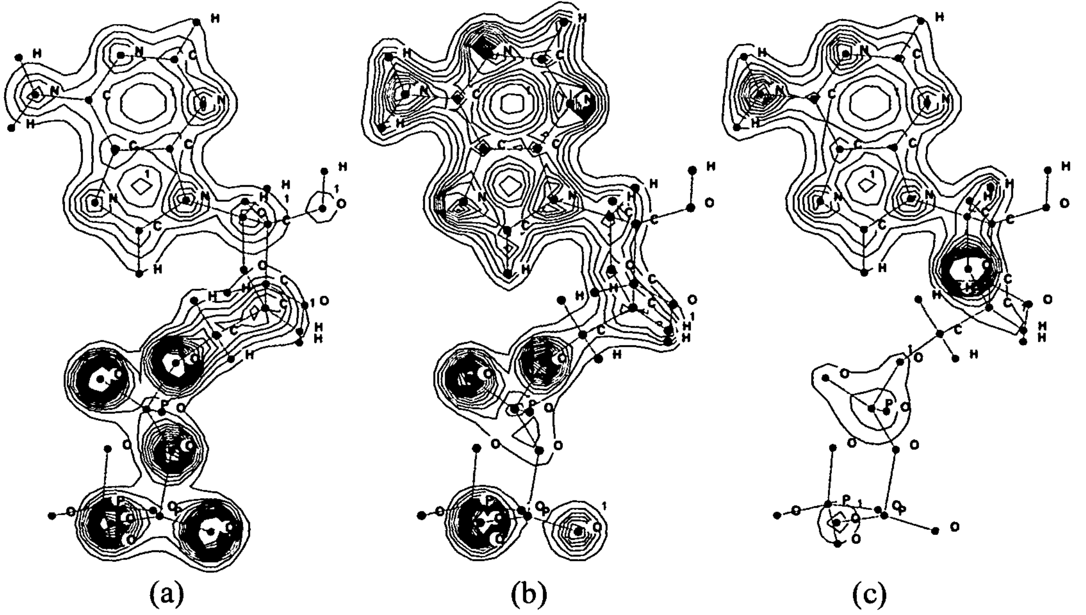


Figure 2-1: Electron density plots parallel to the adenosine ring plane in ATP.
(a) 1.0 Å below, (b) at and (c) 1.0 Å above the ring plane.

of the atoms (see eq. 2-5). The function parameters c and α are determined for each element in the periodic system. The total molecular electron density is then the sum of all atomic functions and can be evaluated for every point **x** in the space:

$$\rho_i(\mathbf{r}) = c_i \cdot e^{\alpha_i \cdot \mathbf{r}} \qquad \text{eq. 2-5}$$

$$\rho(\mathbf{x}) = \sum_{i=1}^{N} \rho_i(\mathbf{r}_{ix}). \qquad \text{eq. 2-6}$$

To make this procedure efficient a cutoff distance can be introduced, so that only those atoms that are in the proximity of a particular point contribute to its electron density. However, this fast approximation of the electron density can only give a qualitative picture of the real situation. For highly accurate solutions, electron structure calculations are necessary.

2.1.4. Solvent-Excluded and Solvent-Accessible Surfaces

Interaction with the solvent (usually water) is of crucial importance for the activities of biomolecules. The stability, reactivity and structural conformation of proteins and protein complexes is often influenced by effects that involve – directly or indirectly – water molecules. E.g. the stability of a protein/protein complex may be determined by the number of apolar amino acid residues that are hidden from the solvent upon binding, or an inhibitor molecule has to compete with water molecules that occupy the active site. It is therefore extremely important to identify the regions in the molecule that are exposed to or hidden from the solvent.

The surface definitions we know so far do not provide us with this information, because they consider only the volumetric extension and size of the molecule itself and do not take any other interacting particle into account. Neither the van der Waals nor the isodensity surface can tell us if, for example, the small and narrow entrance of a deep cavity can be penetrated by water molecules. The general goal is thus to determine the volume of the molecule that is not available (excluded) for solvent molecules. The border of this volume would then be the molecular surface that is accessible to the solvent.

The general idea behind the solution to this problem is described as follows: A solvent particle is represented by a probe – a sphere of the size of the solvent. This probe is then rolled over the van der Waals surface of the molecule. Lee and Richards [80] defined the solvent accessible surface as the trace of the center of the sphere. This is obtained by simply extending the van der Waals radii of all atoms by the radius of the probe and assembling the surface in a similar way as the van der Waals surface. The disadvantage is that the surface is not smooth and does not represent the real interface between the molecule and the solvent.

A better interpretation of the probe-sphere principle is the solvent excluded surface (aka *molecular surface*) that considers the interface between the probe and the molecule: Not the trace of the center of the probe surface but rather the contact points between the molecule and the probe are combined to form the surface. The surface is thus divided into *contact surfaces* which consist of exposed van der Waals spheres, and *reentrant surfaces* that are formed when the probe is in contact with more than one atom at the same time. The volume circumscribed by this surface is the real solvent excluded volume. This kind of surface was made popular by Connolly's MS program [33].

The advantage of the solvent excluded surface is that it combines the benefits of both the van der Waals and isodensity surfaces. Since it is based on the hard-sphere

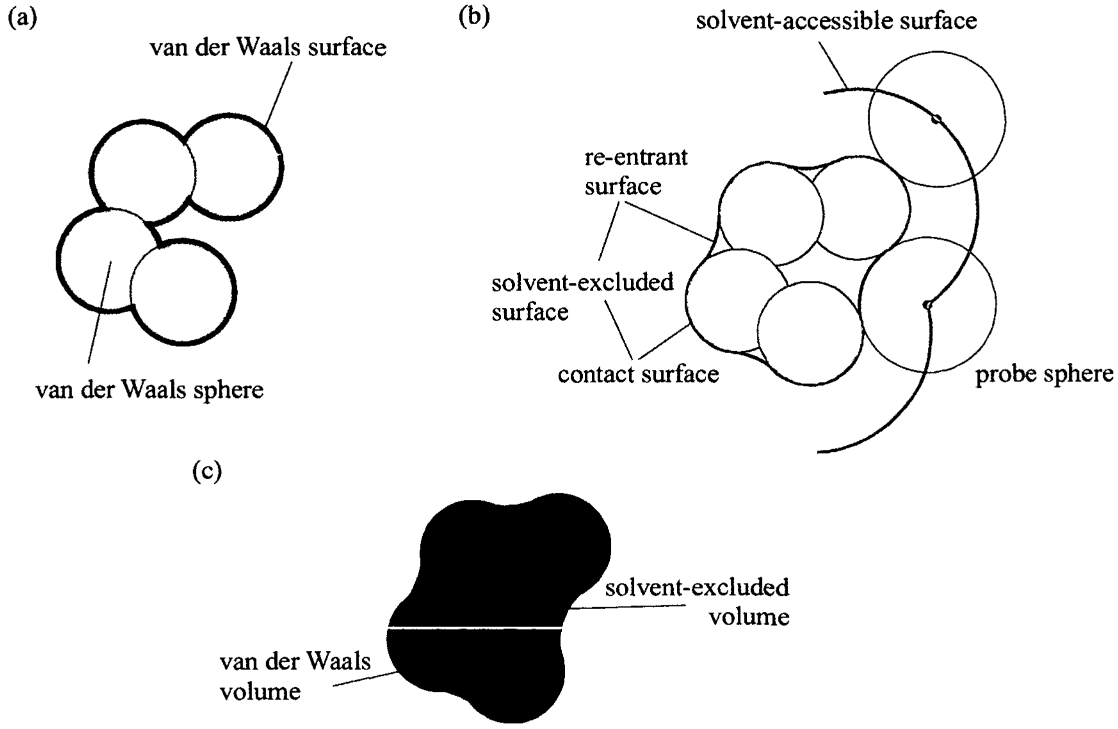


Figure 2-2: Creation of solvent accessible and solvent excluded surfaces.
(a) van der Waals surface (b) creation of the solvent accessible and solvent excluded surface and (c) comparison of the van der Waals and solvent excluded volume. Picture (b) is redrawn from [79].

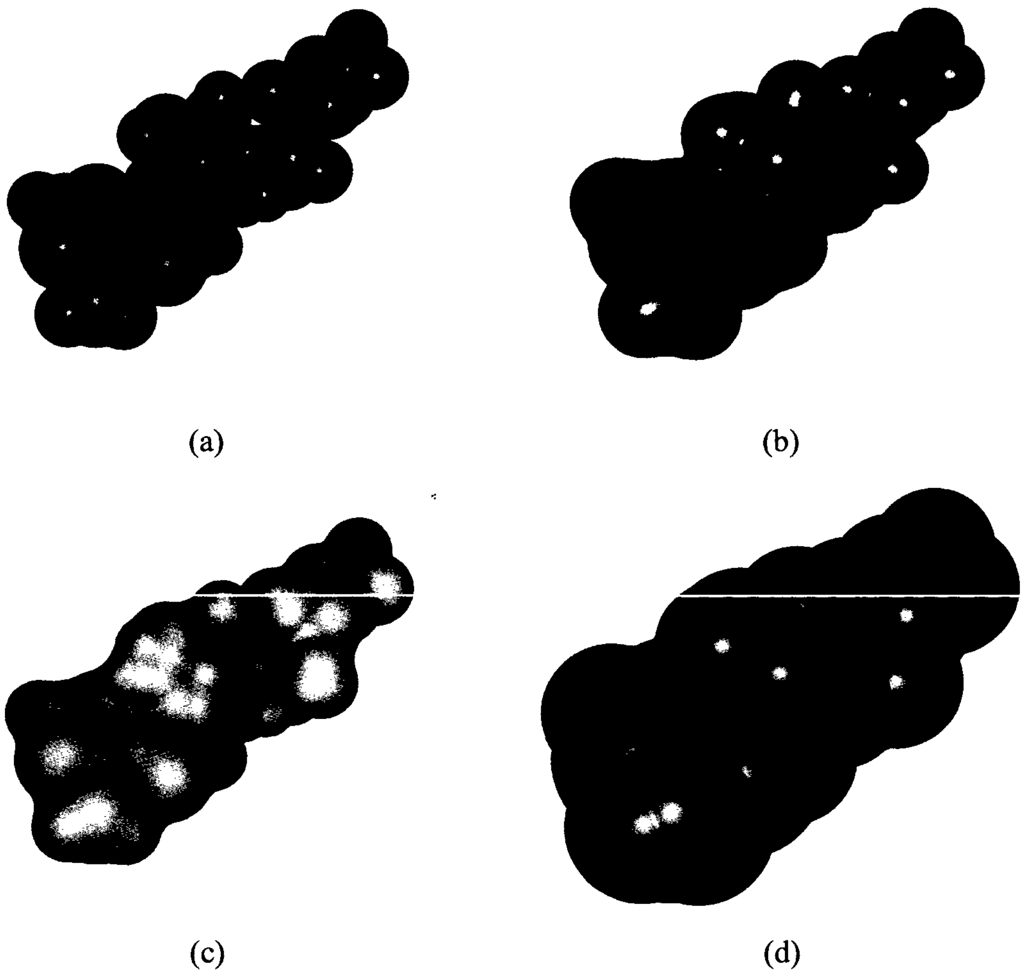
model it can be calculated quickly also for very large systems. It is smooth which makes it possible to calculate curvatures for every point on the surface and it comprises a model that provides more information than the simple 3D hard-sphere arrangement of the van der Waals Surface. Therefore the solvent excluded surfaces have not only become a valuable tool for the calculation and prediction of certain molecular properties but also a popular instrument for the visualization of large proteins and complex systems (see Figure 1-1 on p. 2).

2.1.5. Representation of Molecular Surfaces

Molecular surfaces are very complex geometric objects and in general cannot be assembled from a simple set of sufficiently large building blocks. Depending on the type of the surface different forms of representations are possible:

Analytical description. In special cases it is possible to describe a molecular surface in a closed form: Van der Waals surfaces can be represented by a set of intersecting spheres and solvent excluded surfaces consist of intersecting spheres and torii. The advantage of analytical descriptions is their infinite accuracy and relative compact representation. A disadvantage is the difficulty to extract arbitrary patches from the surface and the calculation of the crossings between the different pieces.

Grid representation. The space around a molecule can be divided into small volumetric elements. Either the centers or the corners of these elements form a 3D grid. Such grids are commonly used when a 3D distribution or property has to be described (e.g. in the finite difference solution of the Poisson-Boltzmann equations or a 3D electron density map). In the same manner it is possible to classify the points on a grid into those which are inside the molecular surface and those which are outside. If a skin, a surface with 3D extension, is considered it is possible to use a three-class model that



(a) (b)

(c) (d)

Figure 2-3: Different molecular surfaces of ATP.
(a) Van der Waals surface, (b) isodensity surface (at level 0.03), (c) Connolly, or Molecular surface and (d) the solvent accessible surface. (scale and orientation of all four pictures is the same)

denotes if a point is outside, on or inside the molecular surface [72]. Unfortunately a grid has a fixed resolution and scales with the third power of the size of the molecule.

Triangulated mesh. Triangulated meshes are widely used in the fields of computer vision and computer-aided design (CAD). They consist of distinct points on the surface which are grouped into a mesh of triangles to form a continuous surface. Unlike the grid, the mesh does not consider the complete space around an object but has also a finite resolution, although by using a large number of small enough triangles it is possible to generate meshes that are comparable in accuracy to analytical surface descriptions. Furthermore these meshes enable the calculation of any surface property by triangular interpolation. Triangle meshes are a suitable way to represent any arbitrary shape by a set of simple graphic primitives, but triangulation is sometimes difficult and not unique.

2.2. Molecular Surface Properties

The different types of molecular surfaces, described in the section above, represent a well defined interface between the molecule and the rest of the system, but they do not provide any additional information about the physicochemical character of that boundary. However, in molecular modeling it is often of great interest to know more about the molecular potentials and properties at the position of the surface points. It is also very useful to calculate some characteristic properties of the surface itself, namely curvatures or surface normals. Many different methods are available in the literature, which enable the mapping of molecular or surface properties onto the surface elements (points and triangles).

2.2.1. Molecular Potentials

In computational chemistry a molecular potential is usually a scalar property that changes with the distance to some distinct feature points. Molecular potentials are in general analytically defined for every point outside and sometimes even inside the molecule. Commonly used and well understood are the electrostatic and lipophilic potential or the hydrogen donor/acceptor density and mapping them onto surface points is straightforward.

Electrostatic potential (ESP). The ESP describes the potential energy of a unit charge in a field of one or several point charges and as such it is a potential in the strict physical sense. In a molecule the electron cloud around each atom has a density that is different from that of the isolated atom due to electron donating and withdrawing groups. Thus every atom can be considered to have a partial charge that reflects the difference between the molecular and isolated environment. The electrostatic field that is built by these charges has an important contribution to the properties and reactivities of the molecule.

Although the ESP for every point on the surface ($\mathbf{p}_i$) can be calculated by means of the electronic wave function, it is more convenient to calculate it classically by Coulomb's law if appropriate atomic charges, q_i, are available (eq. 2-7, with $\mathbf{r}_j$ denoting the position of the atoms). The best approximation is achieved if atomic point charges are used that were fitted to reproduce the real electrostatic potential, calculated from the wave-function. In the present work charges were calculated by the semi-empirical program MOPAC [40] on the AM1 level.

$$ESP(\mathbf{p}_i) = \sum_{j=1}^{N} \frac{q_j}{\left\| \mathbf{p}_i - \mathbf{r}_j \right\|}. \qquad \text{eq. 2-7}$$

Lipophilic Potential (LP). The hydrophobic effect plays an important role in drug-receptor interactions. Diffusion through membranes, solubility of potential drug candidates or propagation within the cellular system are all influenced by the affinity of a molecule to either polar or apolar environments. While not a molecular property itself, lipophilicity can be described empirically by, for example, the n-octanol/water partition coefficient (*logP*). This value is very important for the estimation of many pharmacokinetic properties, but it is difficult to measure. Therefore Ghose and Crippen [54] assembled a table of fragmental *logP* values to calculate this property.

Using these tables we can assign a fragmental lipophilicity value for each atom, f_i, and assign a "lipophilic potential", $LP_{HM}(\mathbf{v}_i)$, to every point $\mathbf{p}_i$ on the surface similar to the ESP [59]:

$$LP_{HM}(\mathbf{p}_i)=\frac{\sum_j^N f_j\cdot g(|\mathbf{p}_i-\mathbf{r}_j|)}{\sum_j^N g(|\mathbf{p}_i-\mathbf{r}_j|)}\ \text{with}\ g(x)=\frac{e^{-C_1C_2}+1}{e^{C_1(x-C_2)}+1} \qquad \text{eq. 2-8}$$

where $\mathbf{r}_j$ is the position of atom j, C_1 and C_2 are experimental constants. Note that LP, in contrast to ESP, is not a potential in the strict physical sense.

Hydrogen Donor/Acceptor Density. The location and density of hydrogen bond acceptor and donor sites is important for the investigation and analysis of proteins and ligand/protein interactions. The concept of hydrogen donor and/or acceptor densities as introduced by Matthias Keil in his PhD thesis [73] is a suitable instrument for the visualization of their distribution on molecular surfaces: For every surface point $\mathbf{p}_i$ a sphere with a given cutoff radius is defined and the number of hydrogen acceptors and/or donors n_{ad} on the molecular surface inside this sphere are counted. This number is divided by the surface area A enclosed by the sphere. Hydrogen donors or acceptors at the border of the cutoff sphere are only counted by the surface part that is located inside the sphere:

$$\rho_{ad}(\mathbf{p}_i)=\frac{\sum_j^{A/D} n_{ad}(j)}{A}\ \text{with}\ n_{ad}(j)=\begin{cases}1 & \text{if } site(j)\in sphere\\ 0\le n\le 1 & \text{if } part\ of\ site(j)\in sphere\end{cases}. \qquad \text{eq. 2-9}$$

Besides this approach there exist other methods to describe the distribution of hydrogen bond acceptors and donors over the molecule or molecular surface (Raevsky et. al. [110;111] or Exner et. al. [48]).

2.2.2. Atomic Properties

In addition to potentials which are usually a feature of the complete molecule, atomic properties can also be of certain interest in some situations. Especially molecular graphics packages like Sybyl [2], VMD [66] or the SWISS PDB [57] use the mapping of atomic properties to display information about the molecular configuration on the surface.

Mapping of these properties onto the surface points is not as straightforward as for the molecular potentials because the atomic properties are only defined for the positions of the atoms and not for all points in space. The usual strategy is to determine the nearest atom for each point on the surface and assign the value of that atom's property to the point. This is a simple technique that has the drawback that the final property distribution on the surface is not smooth. A smooth property distribution can be achieved by means of interpolation techniques, or by the construction of a molecular potential based on that specific atomic property according to the approach used for the lipophilic potential above.

In general different kinds of atomic properties can be mapped onto the surface by one of these methods. Among the most common are the residue number in the sequence, crystallographic B-factors, a color coded residue type, secondary structure types and the partial charge.

2.2.3. Surface Characteristics

Every surface – not only a molecular surface – has certain geometric properties that describe the local shape of the object around a distinct point. These are the different local curvatures and the surface normal. Together with the coordinates of the surface points these quantities provide a full description of an arbitrary 3D surface.

Surface Normals. If we look onto a part of a large surface object that is represented by a triangulated mesh we cannot decide a priori which face of a triangle marks the outside and which the inside. This is a considerable problem in surface visualization, comparison or even creation. When a surface is built an outside and an inside direction has to be defined, according to the particular problem. For the triangles, this definition can be stored in the sequence order of the edge points, so that if viewed from the outside, the three points are arranged in counter-clock-wise order.

Points on the other hand do not have an inside and an outside face, but it is nevertheless necessary to define a vector for each position that indicates the direction away from the surface into the surrounding system. This direction can be defined as the normal vector of the tangent plane to the surface at that particular point with the base at the position of the point and the tip pointing outwards. For each triangle around this point, the tangent planes are trivial, and the normal vectors can be calculated by the cross product

$$\mathbf{n}(t) = -\mathbf{c}_t \times \mathbf{b}_t \qquad \text{eq. 2-10}$$

of the negative of one side vector ($\mathbf{c}_t$) with the vector of its clockwise neighbor ($\mathbf{b}_t$). Surface point normals can then be calculated as the average of the face normals of all triangles adjacent to this point:

$$\mathbf{n}(\mathbf{p}) = \sum_{t=1}^{triangles} -\mathbf{c}_t \times \mathbf{b}_t\,. \qquad \text{eq. 2-11}$$

Normal vectors are usually set to unit length.

Canonical Curvatures. The geometric interpretation of the second derivative of a function is the curvature of its graph. In 3D space a surface object can be expressed or approximated by a function in two variables ($\mathrm{S}_p(u,v)$). The second order derivative of such a function is the Hessian matrix **H** (eq. 2-12).

$$\mathbf{H} = \begin{bmatrix} \dfrac{\partial^2 \mathrm{S}_p(u,v)}{\partial u^2} & \dfrac{\partial^2 \mathrm{S}_p(u,v)}{\partial u \partial v} \\ \dfrac{\partial^2 \mathrm{S}_p(u,v)}{\partial v \partial u} & \dfrac{\partial^2 \mathrm{S}_p(u,v)}{\partial v^2} \end{bmatrix} \qquad \text{eq. 2-12}$$

To accurately describe the shape of the surface we can define canonical curvatures for each point on the surface: A second-order surface (paraboloid) is fitted in a least squares sense to the point and its neighbors within a curvature cut-off range c_{CR}. This paraboloid is the parametrical approximation $\mathrm{S}_p(u,v)$ of the surface around the point p, where u and v are parameters along the principal axes of the paraboloid. The first and second canonical curvatures (cc_1,cc_2) are then obtained as the first and second eigenvalue of the Hessian matrix **H** respectively [141]:

$$\mathbf{H} \cdot \mathbf{d}_1 = cc_1 \cdot \mathbf{d}_1 \text{ and } \mathbf{H} \cdot \mathbf{d}_2 = cc_2 \cdot \mathbf{d}_2 \qquad \text{eq. 2-13}$$

where $\mathbf{d}_{1/2}$ are the directions of the canonical curvatures.

Surface Topology Index (STI). The two canonical curvatures (cc_1, cc_2) cannot be used if an univariate representation of the local curvature is needed. In this case the surface topography index (*STI*) of the MOLCAD [23] program is appropriate (eq. 2-14). Other univariate measures of the local curvature are the mean curvature as described by Desbrun et. al. [39] or the Gaussian curvature (*cg*). The latter is the deviation of the sum of the triangle angles (α_i) at the point from 2π (eq. 2-15).

$$STI = \frac{cc_1 - cc_2}{cc_1} \quad \text{if } cc_1>0 \text{ and } cc_2>0 \text{ or if } (cc_1>0 \text{ and } cc_2\leq 0) \text{ and } |cc_1|>|cc_2|$$

$$STI = \frac{cc_1 + 3\cdot cc_2}{cc_2} \quad \text{if } cc_1\leq 0 \text{ and } cc_2<0 \text{ or if } (cc_1>0 \text{ and } cc_2\leq 0) \text{ and } |cc_1|\leq|cc_2|. \qquad \text{eq. 2-14}$$

$$cg(p) = 2\pi - \sum_{i=1}^{Triangles} \alpha_i \qquad \text{eq. 2-15}$$

2.3. Feature Radius and Auto Correlation

Every surface can be characterized not only by its shape and properties but also by a set of distinct features on it. Surface features are locations on the surface where either the shape or a mapped property belongs to a predefined class. Convex, concave, electrostatic positive or hydrophobic are common feature classes. The difference between the property values at each point within a feature should thus be much smaller than the difference between points of different features.

Features are a form of classification. They can be used to divide a surface into patches of approximately one feature [60] or it may be useful to know how many features are covered by patches of a certain size on average over the surface. The latter can be expressed in terms of the mean feature size or radius which is a characteristic length for a specific surface property. For the calculation of the feature radius one can take the autocorrelation function, as defined by Wagner et. al., who used a spatial autocorrelation of molecular surface properties as molecular descriptors for QSAR calculations [132].

An autocorrelation function AF(*d*) describes the average of the correlations of all property values that are separated by a distance *d*:

$$\text{AF}(d) = \frac{1}{N}\sum_{ij}^{N} p_i p_j \,. \qquad \text{eq. 2-16}$$

where *N* is the number of property (p_i, p_j) values that are separated by the distance *d*. The properties have to be autoscaled to zero average and unit variance in order to give valid results. For molecular surfaces the p_i and p_j are properties at surface points **i** and **j** and the autocorrelation function must be evaluated for ranges of distances, because of the discrete character of the surface points.

If a triangulated mesh represents the surface each point is surrounded by shells of neighbors that are separated by one, two or more edges. We can now apply eq. 2-16 to all possible paths from length 1 up to a length of n_{max} edges. This will give us the autocorrelation function for a hypothetical shift of all points into their first, second or nth shell. To transform the function from the shell into the distance domain we use the average edge length based on the fact that the value of the autocorrelation function is an average *per se*.

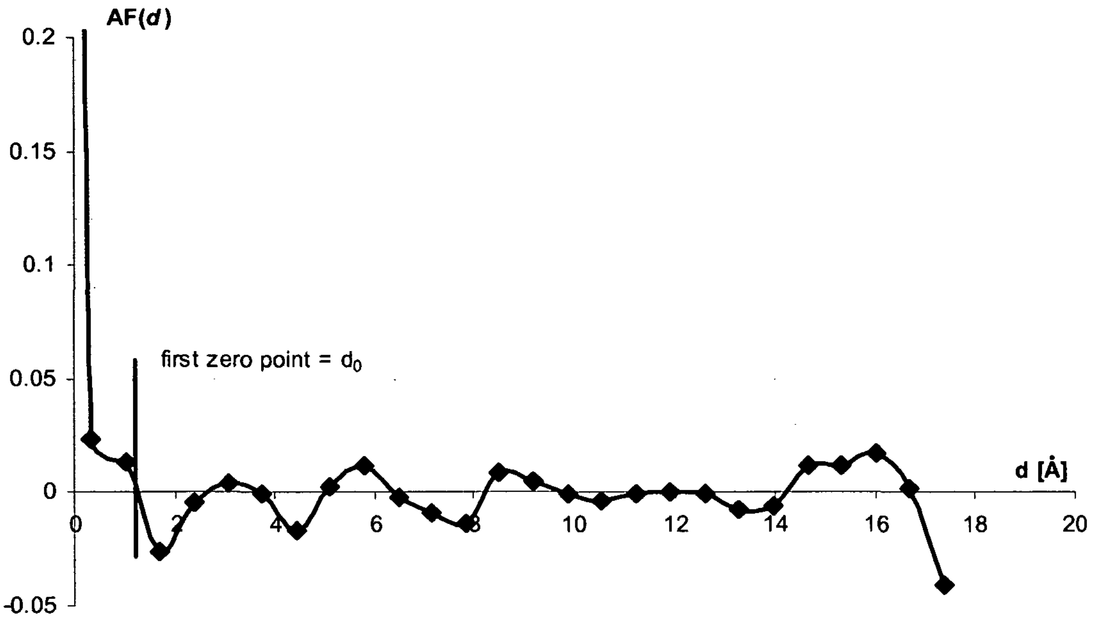


Figure 2-4: Surface autocorrelation function.

Considering that a surface property varies continuously (there are no sharp "peaks" or "edges"), one can expect that points in the immediate neighborhood have similar values and AF(*d*) is thus positive at small distances. Moving farther away the chance that we encounter a region with a completely different property value increases and correspondingly AF(*d*) tends to zero. The distance d_0 where AF(*d*) becomes zero for the first time is taken as the average distance from any point within a specific feature to its border – that is the radius of the feature (Figure 2-4).

In practice the correlation function may not cross the abscissa but come only close to it because of the averaging that is implicit in the autocorrelation. In these cases a statistical t-test was used to check if the AF(*d*) is significant from zero or not.

2.4. Fuzzy Logic

The major task of this work was to detect similarities between molecular surfaces and properties on molecular surfaces. These entities and features are never exactly the same in practice except for an identity comparison. Thus it was necessary to use the right methodology to find similar but not identical properties. Fuzzy logic and harmonic shape images, described below, provide the means for the flexible comparison of molecular surfaces and their properties.

Chemists often use words like “highly negative”, “strongly hydrophobic” or “neutral” to describe the chemical nature of molecular surface regions. These qualitative terms are often accurate enough to distinguish between similarity and dissimilarity among different species in personal discussions or research publications. Unfortunately for computations quality is much more difficult to handle than quantity because classical set theory built upon Boolean logic is restricted to “no membership” or “complete membership” of an object to a specific class. Crisp borders and decision rules are therefore needed for common classification methods. Fuzzy logic, introduced by Lotfi A. Zadeh in 1965 [142], provides a solution for that problem.

A fuzzy set *A*, in contrast to its classical counterpart, does not strictly distinguish between members and non-members (0 or 1) but defines a membership function $\mu_a(x)$

over the definition space (X) that specifies how strongly a value belongs to the set. The value of the membership function is usually normalized to $0 \le \mu_a(x) \le 1$:

$$A = \{(x, \mu_a(x)) | x \in X\}. \qquad \text{eq. 2-17}$$

If we interpret a fuzzy set as a qualitative term like "highly negative" the value of $\mu_a(x)$ defines how well x is described by the term. A linguistic variable LV is a group of fuzzy sets ($A_1 \ldots A_n$) with overlapping membership functions each representing a linguistic term. Therefore it is possible to classify values of x by a scale of terms (e.g. negative, neutral, positive). A linguistic variable LV is defined as

$$LV = \{A_1, A_2 \ldots A_5\} \text{ or}$$
$$LV = \{(x, \mu_1(x)), (x, \mu_2(x)), \ldots, (x, \mu_n(x)) | x \in X\} \qquad \text{eq. 2-18}$$

where $\mu_i(x)$ is the membership function of the i^th fuzzy set (see also Figure 2-5).

Based on these variables Heiden and Brickmann [59] introduced a partitioning function that transforms the qualitative discrimination into a crisp quantitative dissimilarity function D_{LV}:

$$D_{LV}(x,y) = \sum_{i=1}^{N} \frac{w_i |\mu_i(x) - \mu_i(y)|}{w_i(\mu_i(x) + \mu_i(y))} \qquad \text{eq. 2-19}$$

where x and y are two values of the observed variable, $\mu_i(x)$ is the i^th membership function and w_i is the weight for the fuzzy set i. The range of D_{LV} is between 0 and 1 with 0 indicating identity and 1 complete dissimilarity. This fast and simple discrimination function can thus be used to define a qualitative similarity criterion for a quantitative property value.

Since its invention in the 1960-s fuzzy logic has been utilized in many different fields of computational chemistry and cheminformatics. A good overview of the different applications in chemistry can be found in [6].

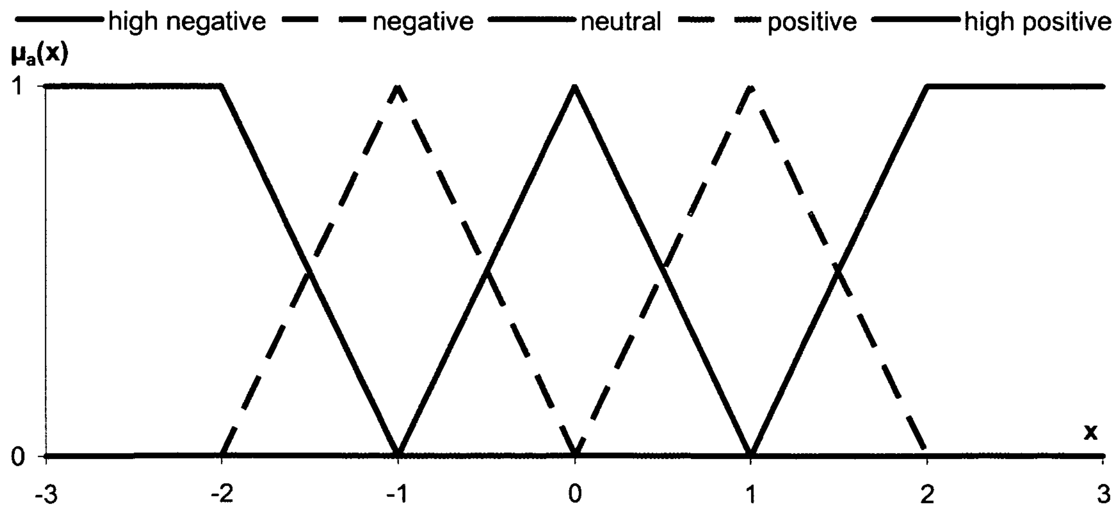


Figure 2-5: Shape of the membership functions in a linguistic variable.
The sets can be used to describe the electrostatic potential on a molecular surface. The variable x represents the autoscaled property values on the surface points and the five classes represent the highly minus, minus, neutral, positive, and highly positive areas of the surface proceeding from left to right.

2.5. Harmonic Images

The preceding section described how properties can be compared in a non-crisp manner by fuzzy logic. This technique is well suited for scalar properties that have been mapped onto the surface. However, the shape of a free molecular surface cannot be expressed by a single scalar value for each point. The topology of the surface and the 3D arrangement of its elements have to be considered as well as the curvature at each point. Moreover the comparison should also remain local, because global similarity comparisons always involve averaging over local features and can thus hide important details. A method for the detection of shape similarity should therefore be able to detect similar features among local regions of the surface, hereinafter called *patches*, and to define correspondences between points on two surfaces based on patch wise similarity.

2.5.1. Concept

Harmonic images [145] provide a methodology to compare patches and to define a relative orientation. They act as 2D representations of 3D surface domains (manifolds) and comparing a complex 3D patch is thus reduced to a rather simple 2D image comparison. The images are generated by using the harmonic mapping method first published by Eells and Sampson [46]. The mapping can be considered as "flattening out" a 3D surface patch P onto a 2D plane D so that an appropriate criterion measuring the distortion is minimized. In the case of harmonic maps and in particular if we consider the approximation introduced by Eck et. al. [44], this minimal distortion criterion can be formulated using a physical analogy:

Let us assume that the edges in the triangulated surface mesh in 3D correspond to ideal springs resting at their equilibrium length. One can assign a "potential energy" level of zero to this undistorted 3D conformation. Mapping onto a flat 2D surface involves stretching and/or shortening of at least some of these imaginary springs and consequently the "potential energy" of the system will increase according to Hooke's law. The harmonic image of the original 3D patch is defined by the arrangement in 2D where this increase in potential energy is minimal.

It can be shown [46] that given a certain boundary there is always a unique harmonic mapping between P and D that constructs a one-to-one correspondence between points on P and vertices on D. Due to this correspondence, any property associated with the points in the original 3D patch can be transferred directly to the corresponding vertices in the 2D harmonic image.

2.5.2. Border Mapping

To obtain comparable harmonic images it is necessary to constrain them to a certain shape, i.e. a unit disk D. This can be achieved by mapping the boundary of the patch directly onto the boundary of the 2D domain. Starting at an arbitrary point at the border of the patch all border vertices of the image are placed at distinct angles of the unit circle using

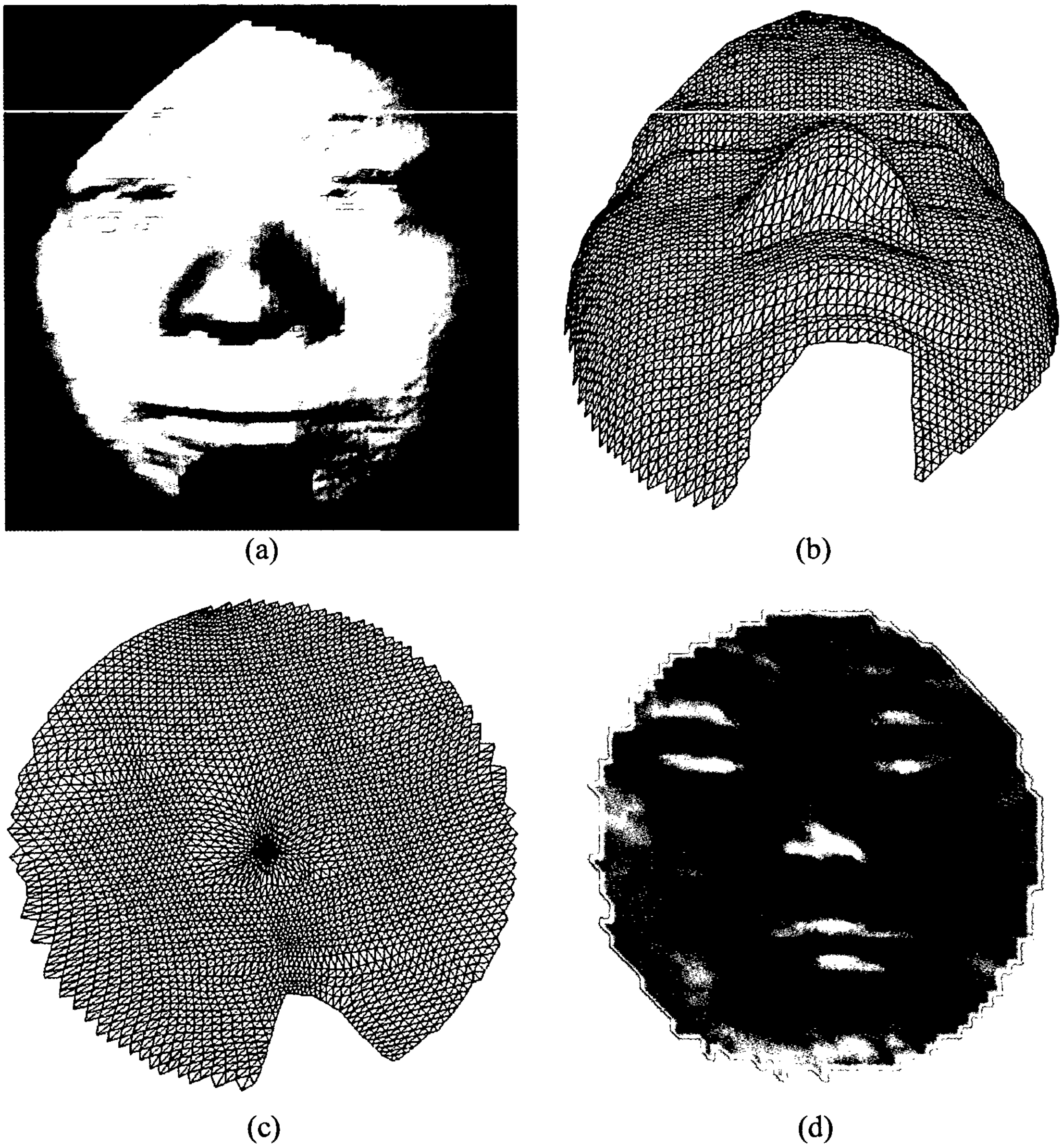
$$\theta_i = \theta_{i-1} + \frac{\alpha_i}{\sum_b^{border} \alpha_b} \cdot 2\pi \quad \text{with } \alpha_i = \angle(\mathbf{p}_i, \mathbf{p}_c, \mathbf{p}_{i-1}). \qquad \text{eq. 2-20}$$

where θ_i and θ_{i-1} are the angle of the actual and previous border vertex, $\mathbf{v}_i$, $\mathbf{v}_{i-1}$ and $\mathbf{v}_c$ are the actual, previous and central points of the patch, α_i is the angle formed by these points and α_b stands for every angle between two border points on the patch.

2.5.3. Interior Mapping

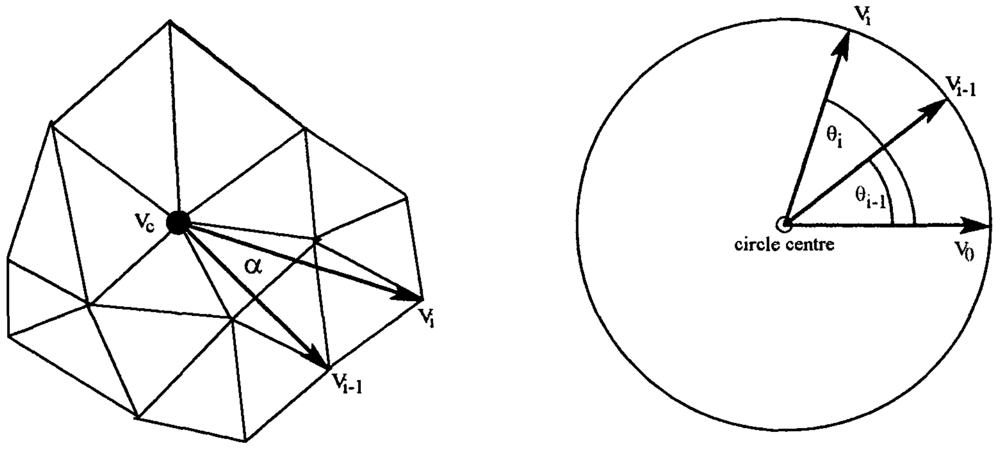
The key step in the generation of the harmonic maps is the solution of an optimization problem. The goal is to minimize the energy function E(ϕ), where ϕ_i and ϕ_j are the mappings of surface points $\mathbf{p}_i$ and $\mathbf{p}_j$ respectively, k_{ij} is the "spring constant" for all possible pairs of points $\mathbf{pp}_{ij}$ and N is the number of surface points in the interior of the patch:

$$\mathrm{E}(\phi)=\frac{1}{2}\sum_{ij}^{N}k_{ij}\cdot\left\|\phi_i-\phi_j\right\|^2 . \qquad \text{eq. 2-21}$$


(a) (b)

(c) (d)

Figure 2-6: A surface patch, its harmonic map and harmonic shape image.
The pictures are redrawn from [144] and show a surface patch of a human face in shaded (a) and wireframe (b) representation. From that patch a harmonic map can be generated (c) which is subsequently resampled into a harmonic shape image (d).



(a) (b)

Figure 2-7: Border mapping.
v_c is the central vertex of the patch v_i and v_{i-1} are adjacent vertices on the border, α is the 3D angle and θ_i and θ_{i-1} are the 2D angles of v_i and v_{i-1} on the map.

In order to find a stationary point of the energy function we have to compute the first derivative of eq. 2-21 with respect to each ϕ_i. The gradient of eq. 2-21 (eq. 2-22) yields eq. 2-23 as the components of a linear equation system.

$$\frac{\partial \mathrm{E}(\phi)}{\partial \phi_i} = \left[\frac{\partial \mathrm{E}(\phi)}{\partial \phi_i^x}, \frac{\partial \mathrm{E}(\phi)}{\partial \phi_i^y}\right] = 0 \text{ for } 1 \le i \le N \text{ and} \qquad \text{eq. 2-22}$$

$$\frac{\partial \mathrm{E}(\phi)}{\partial \phi_i} = \left[\sum_j^N k_{ij} \cdot \left(\phi_i^x - \phi_j^x\right), \sum_j^N k_{ij} \cdot \left(\phi_i^y - \phi_j^y\right)\right]^T . \qquad \text{eq. 2-23}$$

The solutions for the x and y components are independent from each other and can be computed separately. Hence the problem is reduced to the solving of two systems of linear equations. These systems are determined by the "spring constants" k_{ij} that describe whether the imaginary spring connecting the points $\mathbf{p}_i$ and $\mathbf{p}_j$ is stretched easily (k_{ij} is small) or not (k_{ij} is large). In the method, described in this thesis, the spring constants were defined to be inversely proportional to the corresponding edge length in the triangulated mesh of the 3D patch, so that long links between patch vertices could be distorted easily [144]. If the points $\mathbf{p}_i$ and $\mathbf{p}_j$ are not connected by an edge in the triangulated mesh then the constant k_{ij} is set to zero:

$$\frac{\partial \mathrm{E}(\phi)}{\partial \phi_i^a} = \sum_j^N k_{ij} \cdot \left(\phi_i^a - \phi_j^a\right) = 0 . \qquad \text{eq. 2-24}$$

Taking a single row of the linear system, representing the energy function of a distinct mapping ϕ_i eq. 2-24 is the first derivative of the energy function of ϕ_i^a with respect to a component a (either x or y). In each equation the sum over all possible pairs *(i,j)* can be reduced to the sum over all direct neighbors (the one-ring) of ϕ_i:

$$\sum_j^{one-ring} k_{ij} \cdot \left(\phi_i^a - \phi_j^a\right) = 0 . \qquad \text{eq. 2-25}$$

That sum can be split into the sum over all neighbors on the border of the patch (because of the different mapping strategy applied to them) and the sum over all neighbors in the interior region (eq. 2-26). Reordering of the terms by coefficients for ϕ_i^a, ϕ_j^a and ϕ_b^a leads to eq. 2-27, which is suitable for the matrix representation of the equation system (eq. 2-28).

$$\sum_{j}^{interior} k_{ij}\cdot\left(\phi_i^a-\phi_j^a\right)+\sum_{b}^{border} k_{ij}\cdot\left(\phi_i^a-\phi_b^a\right)=0 \qquad \text{eq. 2-26}$$

$$\left(\sum_{j}^{one-ring} k_{ij}\right)\cdot\phi_i^a+\sum_{j}^{interior}-k_{ij}\cdot\phi_j^a=\sum_{b}^{border} k_{ij}\cdot\phi_b^a\,. \qquad \text{eq. 2-27}$$

Hence the systems of linear equations are

$$\mathbf{A}\cdot\phi^x=\mathbf{b}^x \text{ and } \mathbf{A}\cdot\phi^y=\mathbf{b}^y \qquad \text{eq. 2-28}$$

with

$$A_{ij}=\begin{cases}\sum_{l}^{one-ring} k_{ij} & \text{if} \quad i=j\\ -k_{ij} & \text{if} \quad j\in \mathit{one\text{-}ring(i)}\\ 0 & \text{if} \quad j\notin \mathit{one\text{-}ring(i)}\end{cases} \qquad \text{eq. 2-29}$$

$$b_i^a=\begin{cases}0 & \text{if } i \text{ not next to the border}\\ \sum_{b}^{border} k_{ib}\cdot\phi_b^a & \text{otherwise}\end{cases} \qquad \text{eq. 2-30}$$

where **A** is the system matrix defined by the spring constants between all n interior points and $\mathbf{b}^x$ and $\mathbf{b}^y$ are describing the contribution of the border vertices.

2.5.4. Generation and Comparison of Harmonic Shape Images

Generation. Harmonic shape images are a specialization of harmonic images augmented by information about the shape of the original patch. This is achieved by assigning the value of an univariate shape descriptor such as the STI (see section 2.2.3 p. 12) of every point on the 3D patch P to the corresponding vertices on the 2D harmonic image D. As the vertex topology of two harmonic images is almost always different any comparison must be based on a regular grid scheme that is identical for

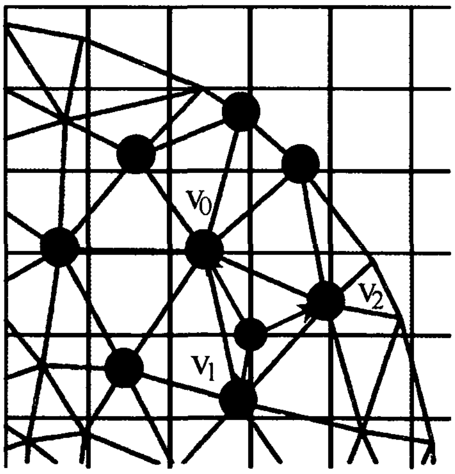


Figure 2-8: Interpolation scheme for the generation of harmonic shape images

both patches.

Hence it is appropriate to replace the original harmonic map with a quadratic $n \times n$ grid where the lateral resolution n is equal to the square root of the number of points n_p in the patch (Figure 2-8). This resampling is done by a triangular interpolation. The triangle beneath every grid point is selected, and the position of the grid point is expressed in its barycentric coordinates, reflecting the influence of each vertex in the triangle to the grid point. Barycentric coordinates for any point within a triangle can be computed with the equations in eq. 2-31 and the interpolated value for the grid point v_G at *(x,y)* is calculated by triangular interpolation with these coordinates:

$$\begin{pmatrix} x_0 & x_1 & x_2 \\ y_0 & y_1 & y_2 \\ 1 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} = \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \qquad \text{eq. 2-31}$$

$$v_G = \alpha \cdot v_0 + \beta \cdot v_1 + \gamma \cdot v_2 \qquad \text{eq. 2-32}$$

where v_0, v_1 and v_2 are the values of the adjacent vertices on the map with the coordinates $x_{0\text{-}2}$ and $y_{0\text{-}2}$ respectively. α, β and γ are the barycentric coordinates of the grid point in the triangle formed by v_0, v_1 and v_2.

The standard resampling scheme by means of a quadratic grid has the disadvantage that only about 75% of the grid points are within the map's range (hence reducing the resolution of the image by approximately 25%). This problem can be solved by a circular grid where all points lie within the unit disk (Figure 2-9). The coordinate transformation was described by Mukundan and Ramakrishnan [101] and can be computed as follows:

$$r = \frac{2\gamma}{N}, \quad \theta = \frac{\pi\xi}{4\gamma} \text{ with } \gamma = \max(|x|,|y|) \qquad \text{eq. 2-33}$$

$$\begin{aligned} |x| = \gamma: &\quad \xi = 2\left(\gamma - x)\frac{y}{|y|} + \frac{xy}{\gamma}\right) \\ |y| = \gamma: &\quad \xi = 2y - \frac{xy}{\gamma} \end{aligned} \qquad \text{eq. 2-34}$$

A circular grid also has a higher symmetry than a rectangular grid which allows a faster computation of the relative rotations. The trade-off is that the points are no longer uniformly distributed, but this has practically no effect on the quality of the results.

The harmonic shape images are stored as vectors of pixels on the circular grid so that each index represents the same grid point in every image of the same resolution.

Comparison. The similarity of two harmonic images can be expressed by the normalized correlation coefficient R of their N-dimensional vectors of pixels $\mathbf{p}$ and $\mathbf{q}$:

$$R = \frac{N \cdot \sum_{i=1}^{N} p_i \cdot q_i - \sum_{i=1}^{N} p_i \cdot \sum_{i=1}^{N} q_i}{\sqrt{\left| N \cdot \sum_{i=1}^{N} p_i^2 - \left(\sum_{i=1}^{N} p_i\right)^2 \right| \cdot \left| N \cdot \sum_{i=1}^{N} q_i^2 - \left(\sum_{i=1}^{N} q_i\right)^2 \right|}}. \qquad \text{eq. 2-35}$$

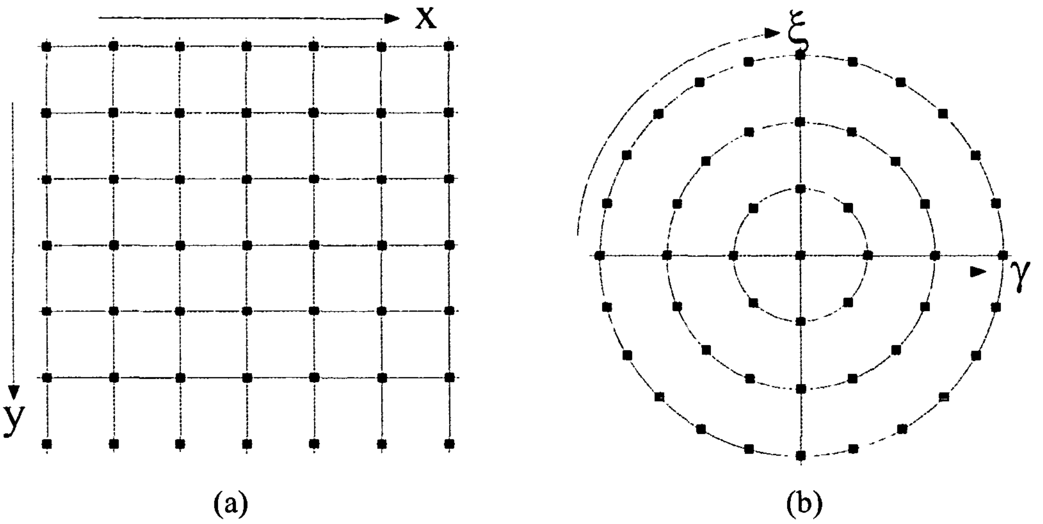


Figure 2-9: Grid transformation.
A rectangular grid (a) can be transformed into a circular oriented raster (b) by the transformation given in eq. 2-33

Because of the arbitrary selection of the first border vertex in the mapping of the patch boundary onto the unit circle (see section 2.5.2 on p. 17), two harmonic images can be rotated against each other. The correlation coefficient is thus a function of the rotation angle θ, and the similarity is defined as the maximum of the correlation function R that is obtained when one image **q** is consecutively rotated against the fixed image **p**:

$$S = \max_{\theta} R(\mathbf{p}(0), \mathbf{q}(\theta)). \qquad \text{eq. 2-36}$$

The first idea in the course of this project was a straight application of the harmonic shape image methodology as proposed by the doctoral thesis of Zhang [144]. Although this approach worked quite well for computer vision problems it did not succeed with molecular surfaces. Zhang used a two step procedure with coarse and fine level searches to detect common features between a template patch and a query surface. The coarse level is an arbitrary sampling of patches uniformly distributed over the query object. In the fine level all points around the best matching patches are used as centers of new patches to find the exact match for the template. The test objects in this work were surfaces of macroscopic items like faces, animals or tools. These objects usually have very distinct but few features and the harmonic images do not change dramatically between two overlapping patches. Molecular surfaces on the other hand have a high feature frequency but the single features are not so significant like a nose in a face. Therefore two overlapping patches can be very different and a coarse level search that is arbitrarily sampling from the molecule's surface will most probably fail in finding the closest matches for a template patch.

Because of this all possible patches on the query surface have to be tested against the template patch which means that a patch is needed for every surface point; but the problem is even more complex. In contrast to the computer vision experts a molecular modeler does not only want to check a single template patch against a query surface but a complete template against a complete query surface. This means that one has to run a harmonic shape image search for every possible patch on one surface against all patches on the other surface. This is a very time consuming operation that scales with the square of the surface sizes and produces a very large amount of result data. Furthermore the results contain a lot of garbage that has to be filtered out. Altogether these problems prohibit a direct usage of Zhang's surface comparison procedure.

In the final procedure that is implemented by SURFCOMP harmonic shape images are still kept as the key elements for shape comparison, but they are applied only to a selected and prefiltered set of patchpairs. However, their comparison remains the time critical step, because for every patchpair the program has to do several resamplings from the map to the circular grid to cover the rotation variance of the harmonic images. To speed up this process a rotation invariant description of the images was tested which was based on Zernike moments [14;143]. These descriptors are following the general moment theorem and are based on radial and angular Zernike polynomials. This technique has been successfully applied to shape analysis and pattern recognition [11;19;62;63;74;87;89]. Because of its rotation invariance the method generates a single representation for each image that can be compared to the moments of other images and can reveal the similarity of the two images and their displacement against each other. Unfortunately this approach failed when applied to molecular surface patches. Although the calculation of the image similarity data was done in a fraction of the time that was needed for the rotation variant approach, the data did neither correlate with the Pearson correlation coefficients nor did it find the correct matches. The reason for this is maybe caused by the less pronounced features of the molecular surfaces which lead to smooth but fuzzy borders between i.e. concave and convex or positive and negative regions. In the literature Zernike moments are usually applied for binary images and applications for gray-scale pictures are rare.

2.6. Maximum Common Subgraph Isomorphism

When one is looking for local similarities between geometric objects such as surfaces sooner or later it will be necessary to combine the similar but local pieces of the puzzle into a complete picture of the global similarity. This is not a trivial task,

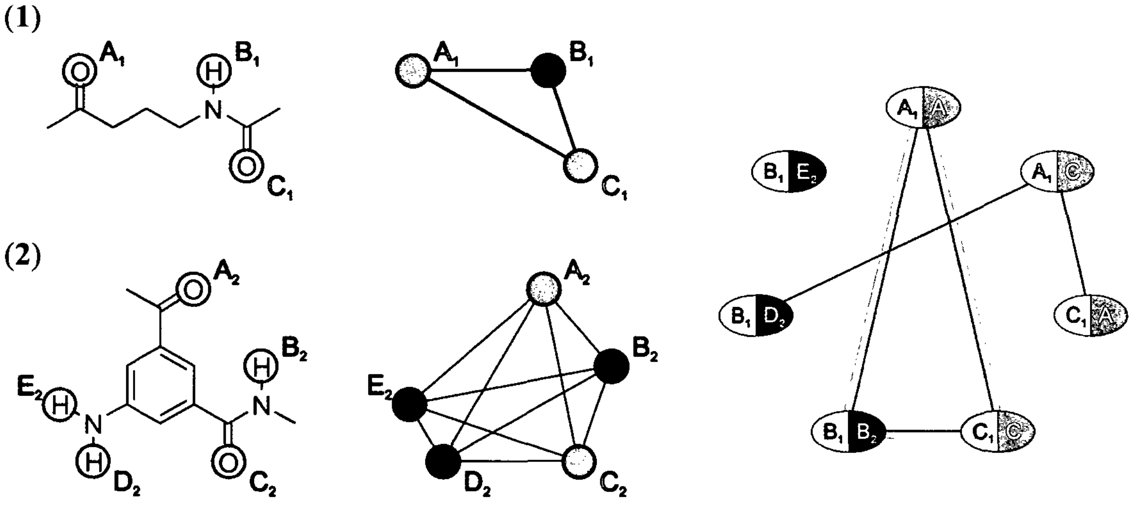


Figure 2-10: Pharmacophore match by maximum common subgraph isomorphism.
The pharmacophoric points on each molecule (a) are transformed into completely connected graphs (b). In the association graph (c) that is formed by the combination of all similar nodes in both graphs only those pairs are connected that have similar edges (c). The triangle ($A_1|A_2$), ($B_1|B_2$) and ($C_1|C_2$) form the only clique in this example and thus represent the match between the two pharmacophores. Example taken from [79]

because not all pieces will fit together and there will be a potentially large number of possible solutions.

One can find some analogies between the combination of matching pharmacophoric points or atoms and the assembling of local surface similarities to a picture of the total resemblance [25;91]. In both cases corresponding features – with a specific location in space – are tested against other local matches to decide whether they represent similar geometrical arrangements. The problem is thus transformed into the detection of similar constellations of points in 3D space which can be solved by means of maximal common subgraph isomorphism. A widely-used method for this is the algorithm of Barrow and Burstall [12] which builds up an association graph followed by clique detection to find the maximum common subgraphs between two query graphs.

Let us consider, for example, a set of pharmacophoric feature points. In this case it is not immediately obvious to see the graph, because usually the points are not connected to each other (Figure 2-10a). However if we consider the steps of the algorithm, as described below, it can be shown that the point sets must be transformed into completely connected graphs (i.e. every point must be connected with every other point in the same set). This indicates that all the points in one set are in a fixed distance to each other which is stored together with the edges (Figure 2-10b).

In a first step one can construct a list that contains all single features of one molecule which are similar or equal to features of the other molecule. Two pairs in this list can only match together if their corresponding features in both molecules are separated by approximately the same distance. This condition holds also for three or more pairs. So eventually only those pairs can contribute to a particular match, which are formed by features that are more or less equidistant to each other on both molecules. If the pairs of similar features are represented by the nodes of a so called association graph an edge can be drawn between all approximately equidistant pairs and the feature sets which form matches between the two molecules can be identified as maximal complete subgraphs or cliques (Figure 2-10c).

The final step of the isomorphism algorithm is thus a clique detection to find all the possible matches explicitly. This is an NP-hard problem [71] and in general we have to resort to approximate solutions. The most commonly used algorithm is due to Bron and Kerbosch [26], an efficient method that uses backtracking and branch-and-bound techniques to perform an exhaustive search for maximal complete subgraphs.

3. Methods

Molecular surfaces are usually represented by triangle meshes containing up to several thousand points. The comparison task is thus to find associations between the point sets of two different surfaces where similarity or complementarity is usually defined by the spatial arrangement of one or more equally defined properties. In this process it is not sufficient to associate every point on surface A with the point on B that gives the best match but it is necessary to consider also the similarities of the corresponding neighbors of A and B. Finally the spatial arrangement of the elements of the detected associations must also be taken into account. The problem of detecting similarities between 3D point sets is well known in cheminformatics. It has been shown earlier that it is equivalent to the maximum common subgraph problem and can be solved efficiently by maximum subgraph isomorphism detection [25;91] (see section 2.6). Unfortunately this is an NP-complete problem [71] which is not a critical limitation for the comparison of small molecular structures with some dozens of atoms, but which makes it inappropriate for the large point sets of complex surface objects. Consequently, if one wants to apply this algorithm to molecular surfaces the number of points has to be reduced and additional information about the chemical and geometrical environment should be represented in a way that is appropriate to dramatically simplify the association graph. In the following sections several heuristics will be discussed that can accomplish this simplification.

The initial point set usually contains a lot of redundant information. The situation around a particular surface element is not really different compared to the environment of its neighbors. But removing all the redundant points does not solve the problem, because it would not be a smooth and accurate representation of the molecule's shape which is needed in the evaluation and refinement process. It is therefore necessary to compare molecular surfaces on at least two different levels of detail: A coarse, non-redundant, representation may be used for the detection of general features that should be matched, and the detection of the correct alignment may be done on a high resolution basis.

Recently, several publications on molecular surface comparison reported successful applications of this idea [37;56]. In particular, Cosgrove et. al. [37] reported a graph-based method that utilizes this two-level approach. On the coarse level, they described the surfaces by patches of the same shape type (convex, concave, saddle shaped, cylindrical and flat). Local geometry parameters are used to decide which patches could overlap and to form an association graph. Matches between the surfaces are then established by clique detection and confirmed by a rigid body alignment at the high resolution level of the corresponding surface points. Their program, called SPAt, gives good results in reasonable time, but they do not consider the chemical environment of points on the surface.

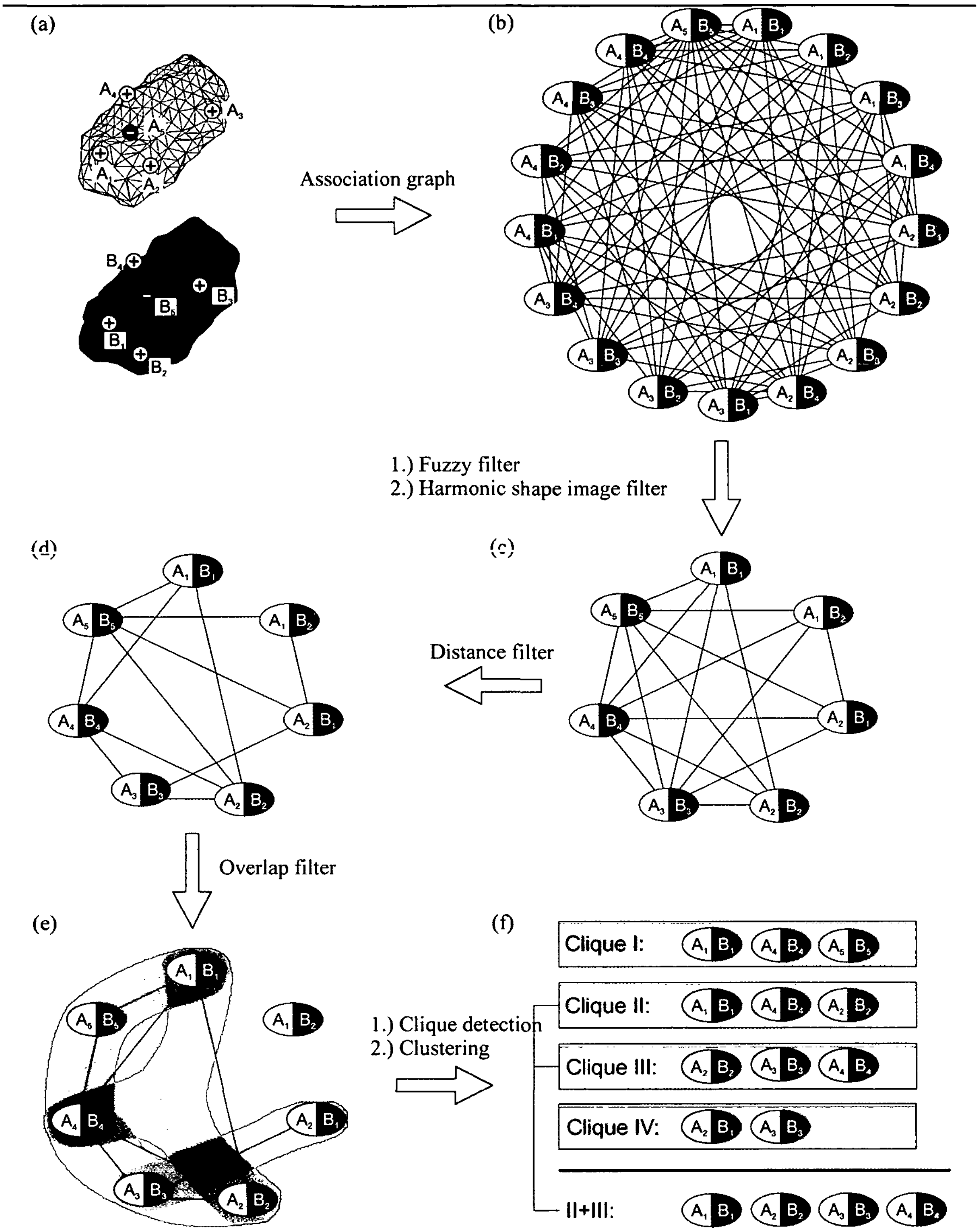


Figure 3-1: Overview of the surface similarity detection algorithm, developed in this thesis.
Starting from two molecular surfaces the critical points are identified (a) and an initial association graph is built (b), which is then further simplified by the fuzzy and harmonic shape image filter (c), the distance filter (d) and the overlap filter (e). From the final association graph the cliques are detected (green, orange, blue and grey regions) and merged (clique II and III in this example) to yield the maximal surface similarity (f).

3.1. General Concept

The general approach, which has been implemented in the SURFCOMP program, is to generate a representation of the surfaces using slightly overlapping circular patces and keep track only of a set of shape critical points (*CP*, coarse level) corresponding to

process step	section[a]	points[b] A	points[b] B	nodes	edges
at the beginning	-	*1131*	*1265*	1.47×10^{6}	2.04×10^{12}
After					
- critical point detection	3.2	27	29	553	274,841
- fuzzy property filter	3.5	24	27	162	17,982
- harmonic shape image filter	3.6	18	25	63	1,260
- distance filter	3.7	18	25	63	359
- overlap filter	3.8	18	25	60	93

Table 3-1: Complexity of the association graph
The number of graph nodes and edges is given for different steps of the filtering process shown for the comparison of 1THL (A) and 4TMN (B).
[a])the section of the text where the step is described
[b])the number of distinct surface points left in the nodes of the association graph

the centers of those patches. The idea of critical points was explored by Connolly's docking algorithm [35] which was later improved by Lin et. al. [86] and Wang [133]. It reduces the number of possible point pairs and associations by several orders of magnitude, so that it is possible to build an initial association graph. This graph is further simplified by several filters that compare the physicochemical properties, surrounding shape and local arrangement of the critical points on both surfaces. (Table 3-1 illustrates the complexity of the association graph at the initial stage and after every step of the algorithm.) In the final graph the similarities are then retrieved by clique detection and rigid body alignments are produced on a point-based (high resolution) level for every match. (see Figure 3-1)

For efficiency reasons SURFCOMP emphasizes the simplification of the association graph which results in a set of smaller cliques that represent only local surface similarities. Therefore, to get a picture of the total similarity between two surfaces, the cliques must be combined to reproduce the complete, global match. For that a hierarchical clustering was used to finally combine those cliques that represent the same geometrical transformation of one molecule onto the other. The final result can be a long list of possible alignments. To provide a faster access to the most promising matches a ranking mechanism was developed and several scoring functions were implemented to sort the results by significance. The alignments can be scored by their size, the root mean square deviation, and the correlations of property values on corresponding points. A ranking is then established by a consensus scoring similar to the methods used in molecular docking.

In a multi step filtering protocol, such as SURFCOMP, several heuristics are used to separate the significant from the insignificant similarities or complementarities. These heuristics are usually controlled by a set of parameters which demand a lot of experience and patience to be tuned properly. The SURFCOMP method needs 7 filtering parameters not including the variables that are involved in the generation of the surfaces and the surface patches. Some of these parameters have proper default values that can be applied to most of the problems, but some other values can have a big influence on the outcome of the calculations. This is not an unsolvable problem and multi step filtering procedures can produce good results (including this thesis), but it should be mentioned that there are alternative methods that have the potential to avoid at least some of these heuristics.

The first heuristic introduced into 3D object comparison is the selection of the coarse and fine representation of the surface. The use of shape critical points is a well understood technique which has been applied to many similarity- and docking-problems but it introduces an arbitrary resolution level that cannot be varied easily. Multi resolution analysis is a computer vision technique that allows the free specification of an object's resolution by means of regular triangulation meshes and wavelet decomposition [45;97]. This approach allows the automatic adjustment of an object's resolution for visualization or storage purposes. The resolution can always be increased or decreased if more or less wavelet coefficients are used. This technique or a similar approach could be used to perform a comparison between two molecular surfaces at a very low resolution, identify the best matching areas, increase the resolution and continue until the match cannot be improved anymore. The iterative procedure would reduce the number of heuristics to the absolute minimum and it would provide an automatic convergence criterion.

The remaining part of this section describes each stage of the method in detail and the steps that are necessary to prepare the input data and evaluate the output. The implementation aspects for each step are given. For the theoretical and algorithmic details the reader is referred to Chapter 2 (Theory). At the end of this chapter a description is given how the application of this method can be extended to the comparison of protein surfaces.

3.2. Definition of Critical Points

The shape of a surface is mainly determined by the location of convex, concave and saddle shaped features. If two objects match, the features of their surfaces have to be similar and should be aligned by the same rigid body transformation. A single feature is usually formed by many surface points which have similar curvature patterns. Hence it is reasonable to take the feature level as the low and the point level as the high resolution for the comparison process. To get an appropriate representation of the surface features a subset of so called shape critical points is extracted from the initial complete set of points. Shape critical points are characteristic points where the properties that define the feature are a maximum.

Shape features can be identified by the signs of the canonical curvatures (cc_1, cc_2; p.

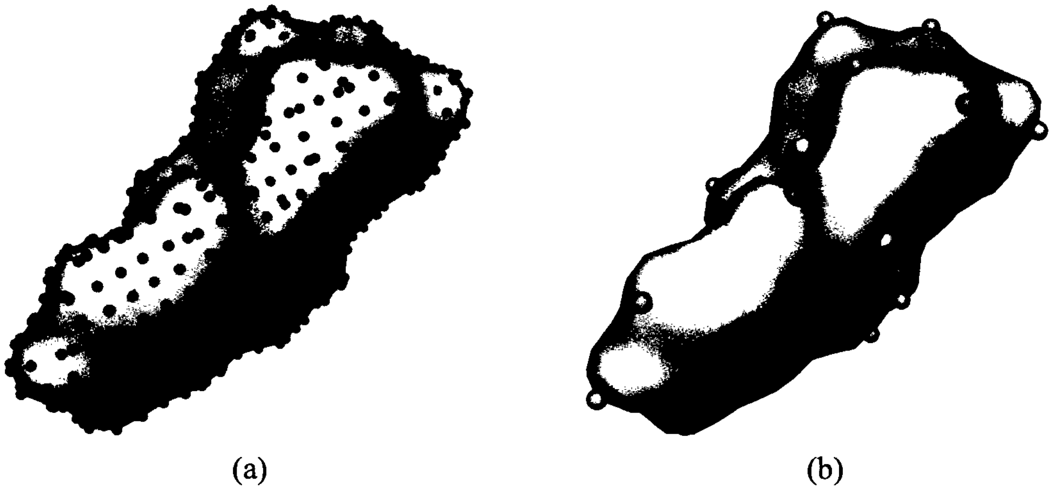

(a) (b)

Figure 3-2: Surface of the thermolysin inhibitor L-valyl-L-tryptophan
(a) all points, (b) distribution of peak (blue) and valley (yellow) critical points over the surface (PDB entry 3TMN).

12): convex regions have two negative, concave two positive and saddle shaped ones display one positive and one negative curvature. Hence two classes of critical points can be defined: A point **p** is a *peak*, if it is a convex point with maximum negative curvature and a *valley*, if it is a concave point with maximum positive curvature in a certain neighborhood N(**p**, r_{cp}) defined by the on-surface radius r_{cp}. This corresponds to a "dip" or "cleft" on the surface (eq. 3-1). To keep the initial set of critical points as small as possible do not consider saddle points are not considered.

$$\begin{aligned} &\mathbf{p} := peak \quad if\ \forall \mathbf{q} \in \mathrm{N}(\mathbf{p}, r_{cp}) \Big| \left|cc_{1\mathbf{p}}\right| > \left|cc_{1\mathbf{q}}\right| \\ &\mathbf{p} := valley \quad if\ \forall \mathbf{q} \in \mathrm{N}(\mathbf{p}, r_{cp}) \Big| \left|cc_{1\mathbf{p}}\right| < \left|cc_{1\mathbf{q}}\right| \end{aligned} \qquad \text{eq. 3-1}$$

At the beginning of a comparison process, before the initial association graph is formed, the peaks and valleys of both surfaces are determinated. The *CP* algorithm investigates every convex or concave point on the surface and adds it to the *peaks* or *valleys* if it meets the appropriate criteria. Figure 3-2b shows the peak and valley critical points of a thermolysin inhibitor molecule. It can be seen that there are many more convex than concave *CP*s. This is due to the fact that most "valleys" are not concave but saddle shaped regions.

3.3. The Association Graph

An initial version of the association graph is formed from the critical points of both surfaces. In this graph the **vertices** correspond to pairs of critical points, $pp_{ij} = (CP_{iA}, CPj_B)$, from the two surfaces that are compared. Hence all the convex and concave critical points of the first surface are paired with the convex and concave *CP*s of the second surface to form the initial set of vertices. This means that at this stage all the critical points with the same curvature attribute are considered similar.

According to the definition of an association graph, **edges** should be drawn between every two pairs that do not have a critical point in common (see Figure 3-1b), but for computational reasons no edges are considered before the application of the distance filter described below.

3.4. Generation of Surface Patches

Since SURFCOMP's coarse level representation of molecular surfaces is the set of their critical points together with their neighborhoods, it is necessary to have a consistent definition of the vicinity of a surface point also known as the *patch*. A patch is a continuous piece of the surface centered on a point **c** that includes all points around that center within a certain *on-surface* distance, henceforth called the patch radius. (An *on-surface* distance of two points is the shortest path between them over the surface, not straight through 3D space).

In this work patches, like surfaces, are represented by triangulated meshes (p. 9). But unlike surfaces which are completely closed objects with no boundaries, a patch has a border that should be defined by an unambiguous sequence of triangle edges. For the harmonic image construction (p. 17) all the points on the border must be passed exactly once before the point where the walk was started is reached again. In such a traversal the iteration from one border point to the other is controlled by the counter clockwise order of the triangle points: In all triangles that contain the active point its successors are examined to find a border point that has not yet been reached. To make this mechanism work, every triangle must not have more than one border edge (because then there would be more than one unvisited border point among the successors, and the

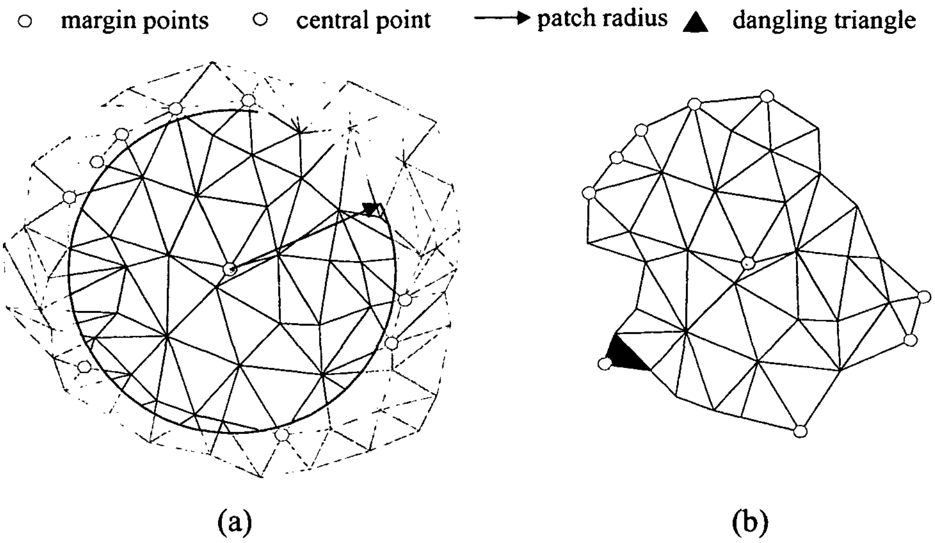


Figure 3-3: Illustration of the patch generation process.
In the first steps the points within and in close contact to the patch radius are selected (a) and extracted from the surface. There may be dangling triangles (b) which have to be removed in the refinement to ensure consistency.

walk could end up in a loop, a shortcut or a dead end). Therefore triangles with two or three edges exposed to the border (dangling triangles) should be removed.

Holes in surface patches present another problem. They are formed if a molecular surface contains "pillar-like" cylindrical areas in the close vicinity to the central point. In this case, only the base of the pillar lies within the patch radius and the upper part or the head is not included. Automatically including all the points in the hole is not appropriate, because if the pillar is a bridge to the rest of the surface, all the other points would be included and the patch would be equal to the complete surface. Fortunately the harmonic shape images are robust with respect to holes and missing parts in patches [144], therefore it was decided not to fill them.

According to the considerations above, the patch around a central point **c** is created as follows (see also Figure 3-3):

1. A subset of the surface points which have an Euclidean distance to **c** that is less than the patch radius is preselected to avoid on-surface distance calculation on the complete surface.
2. The on-surface distances between **c** and all the points in the subset are calculated by the Dijkstra shortest path algorithm [41] with the edge-weights set to the Euclidean distances between neighboring points.
3. All points around **c** within the patch radius are extracted plus any points connected to them that lie within a 5% margin off this radius.
4. Every triangle on the surface that contains three selected points is copied to the patch.
5. To preserve a correct clockwise or counter clockwise walk, all triangles in the patch with more than one border edge (dangling triangles) are removed.
6. If there are any points that do not belong to a triangle, they are removed and step 5. is repeated until consistency is achieved.
7. For each remaining point a reference to the original point in the surface is stored.

3.5. Fuzzy Filter

In order to reduce the complexity of the problem, it is necessary to remove those critical point pairs from the association graph that do not have a similar chemical environment. Each vertex of the graph must thus be checked by a chemical filter to ensure that the corresponding critical points have similar chemical properties. Fuzzy sets and linguistic variables [142] were used to express the similarity between chemical properties mapped onto the surface, and applied a defuzzification function, introduced by Exner et. al. [48] as a similarity measure (p. 15).

According to Figure 2-5 five fuzzy sets were defined for each physicochemical property in the experiments that correspond to common classifications (Table 3-2). An important issue in the application of that technique is the scaling of the compared properties. For every possible value of the property the contribution to each of the predefined fuzzy sets must be specified in advance (see also eq. 2-18 on p. 16). But many quantities that can be mapped onto the molecular surface vary too much to apply fixed boundaries and relations between these sets. Especially the electrostatic potential depends strongly on the total charge of the molecule and consequently it is meaningless to compare surface ESP patterns directly between molecules with different total net charges unless an appropriate normalization is carried out. For instance the absolute difference between the ESP around the adenosine 3-H in ATP^{-4} and ATP^{-3} is about 50 kcal/mol while the relative difference between the ESP over the center of the adenosine 6-ring and the 3-H in the 3 and 4 minus species is only 10 kcal/mol. For a general interpretation of the membership functions it will therefore be necessary to use normalized (mean-centered or autoscaled) surface properties. In the fuzzy filtering autoscaled functions were used. The rationale behind this is that in surface similarity searches, the aim is to find the region on one surface that fits *most likely* to a patch on the other one. Hence the most positive or negative values will fit best to each other regardless of their absolute difference. Furthermore an autoscaled property provides natural classifications for the membership functions.

property	high –	–	neutral	+	high +
ESP	highly negative	negative	neutral	Positive	highly positive
LP	highly hydrophilic	hydrophilic	amphiphilic	Hydrophobic	highly hydrophobic

Table 3-2: Definition of fuzzy qualitative classes
These classes are used as fuzzy sets in the linguistic variables of the fuzzy filter.

The fuzzy filter is the first filtering step in the surface comparison process. It takes every vertex of the initial association graph, and calculates the fuzzy dissimilarity function for a certain chemical property of its points according to eq. 2-19 (p. 16). Every vertex whose points are more dissimilar than a certain fuzzy threshold F is then removed from the graph and its points are considered to be chemically dissimilar. By this filter, the number of the associations can be reduced by approximately 80% (see also Table 3-1).

3.6. Harmonic Shape Image Filter

The fuzzy chemical filtering checks for similar physicochemical properties between both surfaces, but it does not consider the shape of the molecules around the critical points. It is important, however, to consider the surface-patches around the critical points and compare them with each other to establish whether two *CP*s are embedded in similar regions and how their neighborhoods are best oriented relative to each other. In the present surface comparison process harmonic shape images (HSI) [145] (p. 17) provide the methodology to compare the patches and to define a relative orientation between them.

Harmonic shape images compare surface patches by a local shape descriptor mapped onto their points. Several such descriptors were introduced in section 2.2.3. In the present work the surface topology index (STI) [23] was used to compare the shape of two surface patches, but any other scalar shape descriptor could be applied as well. While in general possible, multiple scalar values, such as the canonical curvatures, are not used in the HSI comparison because the Pearson correlation function is susceptible to leverage effects. Such effects can be easily introduced, if two variables do not occupy the same space or if the dissimilarities of one type neutralize the similarities of the other or vice versa.

HSI generation. The transformation of a surface patch into a harmonic shape image is a multi-step process that involves (i) the detection and mapping of the patch border, (ii) the correct mapping of the patch's interior points, and (iii) the sampling of the shape descriptors from the point-based mesh to a regular grid (see also p. 17). Especially the sampling step is computationally intensive and special techniques must be applied to improve its speed.

The detection of the patch's border is done by the patch generation algorithm (section 3.4) when removing the dangling triangles. For the mapping of the border points a continuous, counterclockwise walk along this border will provide us with the correct sequence for the determination of the position angles θ_i on the unit circle (eq. 2-20). The traversal is done by following the triangle edges from one border point to its successor at the boundary. The positions of the border points are obtained in polar form with the radius r=1 and the angles θ_i and must be transformed into Cartesian coordinates for the interior mapping (p. 17).

After the position of the border is fixed, the interior points can be mapped into the unit disk. According to section 2.5.3, these positions are defined by a pair of systems of linear equations (eq. 2-28). The matrices **A** and $\mathbf{b}_x\mathbf{b}_y$ must be assembled according to eq. 2-29 and eq. 2-30 with the spring constants k_{ij} set to

$$\frac{1}{\left\|\mathbf{p}_i - \mathbf{p}_j\right\|} \qquad \text{eq. 3-2}$$

and the actual positions of the border points on the unit circle (p. 17). The fact that the systems use the same coefficient matrix **A** and different constrain vectors $\mathbf{b}_x$ and $\mathbf{b}_y$ for the x and y position, makes it suitable to solve the equations by LU-decomposition [108].

The last step is the generation of the shape image by resampling of the descriptors from the mapped points to a regular grid. To perform the sampling it is necessary to identify the triangle beneath every position at the grid. The actual value for the grid point is then calculated by an interpolation of the triangle's vertices (p. 20, eq. 2-32). The search for the active triangles is of order $O(N^3)$ where N is the number of grid points which is equal to the number of points in the patch. Hence the resampling is the rate determining step of the HSI generation. The process can be accelerated if a geometric hashing algorithm is used to investigate only the triangles in the vicinity of a grid point. To this end the whole image is divided into fields in a way that each field contains approximately 5 triangles. All triangles are assigned to those fields, where at least a part of the triangle is present. Later, during the resampling, only those triangles of the field of the current grid point are considered.

Image comparison. As discussed in section 2.5.4 the images can be rotated against each other. It is thus necessary to perform a full angular scan when determining the similarity between two harmonic shape images. Usually one of the images is fixed, and the other one is rotated by a predefined angular increment δ (usually 2°) and resampled for each rotated position around the circle. To avoid unnecessary sampling the grids for the flexible image are precompiled for every possible angular position and persistently stored with the harmonic map data.

The harmonic shape image filter is invoked for every critical point pair in the association graph that is left after the fuzzy filtering step. In this filter step the patches around the critical points are computed, if they are not yet available, and transformed into the corresponding harmonic images. Then the two images of the *CPs* of the associated point pair are compared as described above and on p. 21. If the detected similarity is better than the shape threshold *R* the point pair passes the filter otherwise it is removed from the association graph.

3.7. Distance Filter

Up to this point only single pairs of critical points (*pp*) have been considered which are represented by the vertices of the association graph. However, the aim is to find groups of *CP* pairs which represent a similarity between the compared surfaces. Thus it is necessary to form edges between the point pairs in the association graph, to identify those which can overlap at the same time.

A simple but effective criterion is the difference of the distances of two point pairs on surface *A* and *B*. Considering two point pairs $pp_1 = (CP_{A1}, CP_{B1})$ and $pp_2 = (CP_{A2}, CP_{B2})$ with the positions of their critical points $\mathbf{p}_{A1}$, $\mathbf{p}_{B1}$ and $\mathbf{p}_{A2}$, $\mathbf{p}_{B2}$, the distances δ_A and δ_B are

$$\begin{aligned} \delta_A &= \left\|\mathbf{p}_{A1} - \mathbf{p}_{A2}\right\| \qquad (A) \\ \delta_B &= \left\|\mathbf{p}_{B1} - \mathbf{p}_{B2}\right\| \qquad (B) \end{aligned} \qquad \text{eq. 3-3}$$

the Euclidean distances between the two critical points on surfaces A and B (see also Figure 3-4). Martin et. al. used the same criterion for the identification of pharmacophore patterns [91].

Two pairs are connected in the association graph only if the distances δ_A and δ_B are within a certain distance tolerance $t \geq |\delta_A - \delta_B|$ and δ_A, δ_B are larger than the minimum

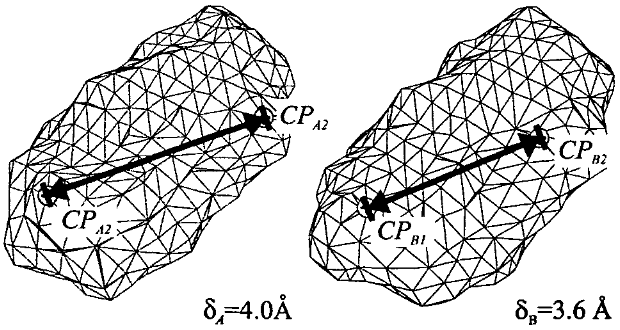


Figure 3-4: Distance filter

distance δ_{min}. The minimum distance is introduced to avoid connections between very close critical point pairs which represent essentially the same regions. It should also be noticed, that no connections must be drawn between critical point pairs that share the same point on either of the two surfaces (i.e. $CP_{A1} \neq CP_{A2}$ and $CP_{B1} \neq CP_{B2}$).

3.8. Overlap Filter

The distance filter checks if two pairs are at an appropriate distance for simultaneous overlap, but harmonic image matching provides additional information about the optimal orientation of each *CP* patch pair. Using this information the number of connections in the association graph can be further reduced.

The idea is to check the simultaneous overlap of both pairs via the relative orientations of the connecting axes on surface *A* and *B*. In Figure 3-5 the axes between the two critical points on each surface are projected onto the harmonic maps of the patches and the closest points on the borders of the patches are determined. α_1, α_2 and β_1, β_2 denote the angles between the optimum orientation (alignment axis) and the closest points to the *CP* axes on surface *A* and *B* respectively. The α and β angles thus describe the heading from one critical point patch to the other with respect to the alignment axis.

The filter computes the heading differences φ_1, φ_2 for both *CP* patch pairs and removes the connection between them, if none of them is within a certain angular tolerance φ_{tol}:

$$\begin{aligned} \varphi_1 &= |\beta_1 - \alpha_1| \\ \varphi_2 &= |\beta_2 - \alpha_2| \end{aligned} \qquad \text{eq. 3-4}$$

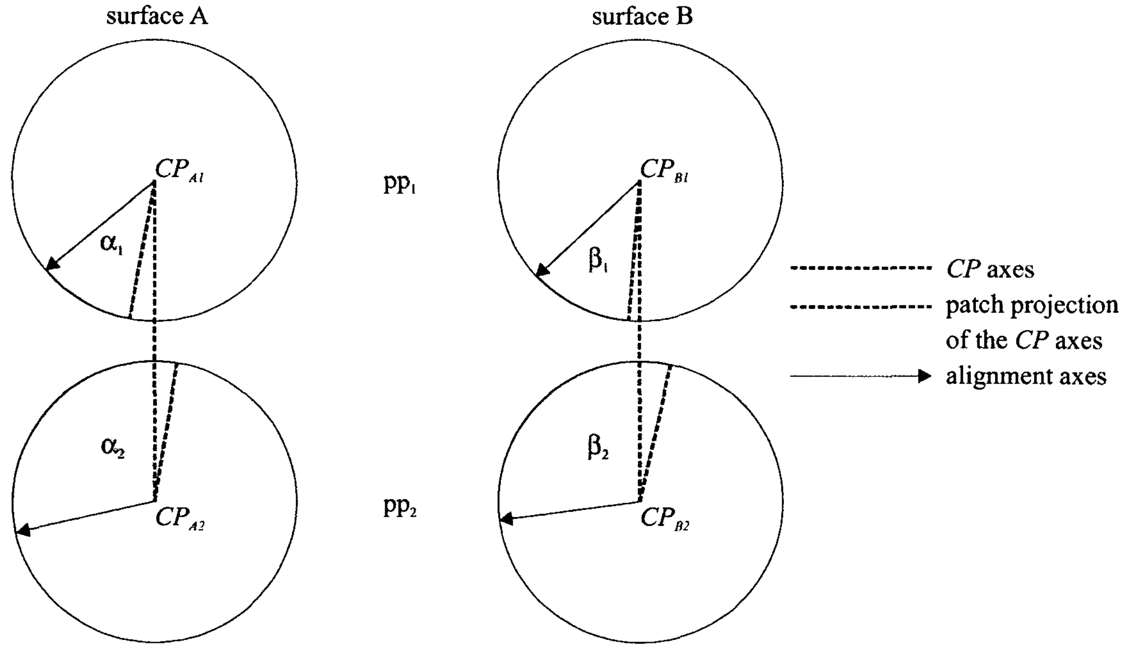


Figure 3-5: Illustration of the overlap filter.
The axes between the two patches on both surfaces (black stippled lines) are projected onto the harmonic map of the surface patch, and the angles between that projections and the axes that define "north" (0°) in the optimal alignment of the patchpairs pp_1 and pp_2 are determined as the bearing from one patch to the other patch on the same surface.

3.9. Clique Detection and Clustering

Having applied all the filters, the size of the association graph is reduced so that it is possible to search for cliques in it. The algorithm of Bron and Kerbosch [26] was used to find all cliques which are present in the association graph. This usually results in a large set of cliques consisting of two to four critical point pairs. They only represent partial similarities and must be combined to get the largest possible local surface alignment between the two molecules. Therefore, these small primary cliques are combined into larger clusters that represent different sets of corresponding points on both surfaces.

For each cluster a rigid body transformation was generated based on all the correspondences detected by the harmonic shape image matching for the patches around the critical points. The transformation matrices **T** are calculated by a least squares fit [93] of the two point sets superimposed over their centers of gravity. The root mean square deviation (RMSD) of this transformation serves as a quality criterion for the cluster:

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N} \left\| \mathbf{T} \cdot \tilde{\mathbf{q}}_i - \bar{\mathbf{p}}_i \right\|^2}{N}} \qquad \text{eq. 3-5}$$

where $\mathbf{p}_i$ is one of the N fixed and $\mathbf{q}_i$ is one of the N transformed corresponding points. From the large set of initial small clusters, those with high RMSD values are eliminated (above 4.0 Å) and the remaining clusters are subject to a stepwise hierarchical-linkage clustering as follows.

For all pairs of clusters in the list that can be combined, the RMS deviations for the transformation of cluster A with the transformation matrix of cluster B and vice versa are calculated; the smaller value (single linkage) is stored as the distance between A and B. Two clusters A, B cannot be combined if a critical point is paired with a different *CP* in A and B. At each step the algorithm takes the two closest clusters and merges them into a new one while updating the distances to the remaining clusters. The new one replaces the merged clusters in the list and the algorithm is repeated until no more clusters can be merged. The result is a set of possible local surface alignments.

Besides single linkage complete and average linkage were examined too, but they did not cause any differences in the quality of the results. Because single-linkage can be implemented more efficiently than complete and average linkage, it was used in all experiments.

3.10. Scoring and Ranking

The hierarchical clustering provides the results of a surface comparison as a tree, where the largest alignments are found in the elements closest to the root and the original cliques are placed in the leaves. This representation can be very useful when one wants to examine how the larger alignments are composed and how strong the different elements of the clusters are related to each other. However, in most of the cases the primary question is which alignment is the best in respect to percentage of the covered surface, quality of the rigid body fit and chemical similarity. Especially when the molecules are large and the comparison produces a number of possible top-level alignments, this task is difficult to do by visual inspection of the alignments even though the pure RMSD value of the rigid body fit provides a good initial guess of the quality of the clusters.

Another problem lies in the nature of the hierarchical clustering: The algorithm subsequently combines two clusters into a new one either until only one cluster is left or the new one cannot be combined with another one because of ambiguous critical point combinations. Thus the top-level clusters often represent poor alignments if two clusters that practically do not fit together are combined because all their *CP* pairs are correct. The consequence is that the best alignments are often placed in the levels beneath the top. It will therefore be necessary to find an alternative ranking that will sort out the promising alignments by a combination of patch-size, geometrical and chemical similarity criteria.

Ranking is a well known problem in molecular docking, where the large lists of possible ligand/receptor conformations must be scored and sorted to simplify and speed up the manual search for the best structure. The usual strategy is to improve the energy score, which is by far the most important criterion, with several other heuristics [29]. Wang and Wang [135] identified three different classes of consensus scoring methodologies: *"rank-by-number"* (eq. 3-6) uses the average scoring value, *"rank-by-rank"* (eq. 3-7) takes the average rank and *"rank-by-vote"* (eq. 3-8) counts how often an entry is sorted into the top x% by each scoring function:

$$r_i = \frac{1}{N}\sum_{j=1}^{N}\mathrm{SF_j}(x_i) \qquad \text{eq. 3-6}$$

$$r_i = \frac{1}{N}\sum_{j=1}^{N}\mathrm{rank}(\mathrm{SF_j}(x_i)) \qquad \text{eq. 3-7}$$

$$r_i = \sum_{j=1}^{N}\mathrm{top}(n,\mathrm{SF_j}(x_i)) \qquad \text{eq. 3-8}$$

where r_i is the rank of the docking result x_i, $\mathrm{SF_j}$ is the j$^{\mathrm{th}}$ scoring function. $\mathrm{rank}(\mathrm{SF_j}(x_i))$ returns the rank of x_i and $\mathrm{top}(n, \mathrm{SF_j}(x_i))$ yields true if x_i is among the top n% according to $\mathrm{SF_j}$ or false otherwise.

A consensus scoring algorithm was implemented in SURFCOMP based on the *rank-by-rank* scheme. The algorithm calculates the average ranks determined by (a) the RMSD value of the rigid body fit, (b) the number of corresponding surface points that build the alignment and (c) the chemical correlation of these points. Thereafter it sorts the results in a way that places the most promising clusters at the top of the list. All clusters are evaluated and ranked independently of their position in the hierarchy.

ad a. The ***RMSD*** value is provided by the clustering algorithm eq. 3-5. The clusters are ranked in ascending order because the quality of a rigid body transformation is inversely proportional to the RMSD value.

ad b. The number of corresponding points (N_{points}) reflects the size of the detected surface similarity. The larger the common surface area the better the cluster is ranked by this quantity. It is somehow complementary to the RMSD because larger point sets are more likely to produce larger RMSD values so combining the RMSD and size of the similarity reflects a kind of trade off between accuracy and size.

ad c. RMSD and the number of corresponding points are responsible for the evaluation of the geometric fit. Besides that a check of the chemical similarity should not be neglected. This can be easily performed by the calculation of a Pearson correlation coefficient $\boldsymbol{R_{chem}}$ (eq. 2-35 on p. 21) between the physicochemical properties of the corresponding points. For the ranking all the clusters are sorted by descending order.

The final ranking value is the weighted average rank over all terms (eq. 3-9) and a sort in ascending order will bring the must promising clusters to the top of the list. Usually the three contributions are weighted equally and the weights are set to 1. Once a good cluster is identified in that list its hierarchical position can be examined to check whether its parent or one of its children may provide a better representation of that particular surface similarity.

$$consensusrank = \frac{1}{3}\left[\mathrm{rank}(RMSD) + \mathrm{rank}(N_{points}) + \mathrm{rank}(R_{chem})\right] \qquad \text{eq. 3-9}$$

3.11. Treatment of Protein Surfaces

The first goal of this project was to establish a surface comparison algorithm for low molecular weight compounds. However, in the course of the studies the ability of the program to compare surfaces of large biomolecules such as proteins was investigated too. The main problem with large molecules is that even at a lower level of resolution the number of points is approximately one order of magnitude greater than for small compounds. Because most of the algorithms in the comparison process scale approximately quadratically with the number of surface points, this implies a massive increase in computational time, memory and the number of candidate alignments that have to be evaluated.

In molecular modeling and drug discovery the functionally most interesting part of a protein surface is where a substrate is converted catalytically (active sites of enzymes), a cofactor is bound or a signal molecule is recognized. Fortunately, such functional sites

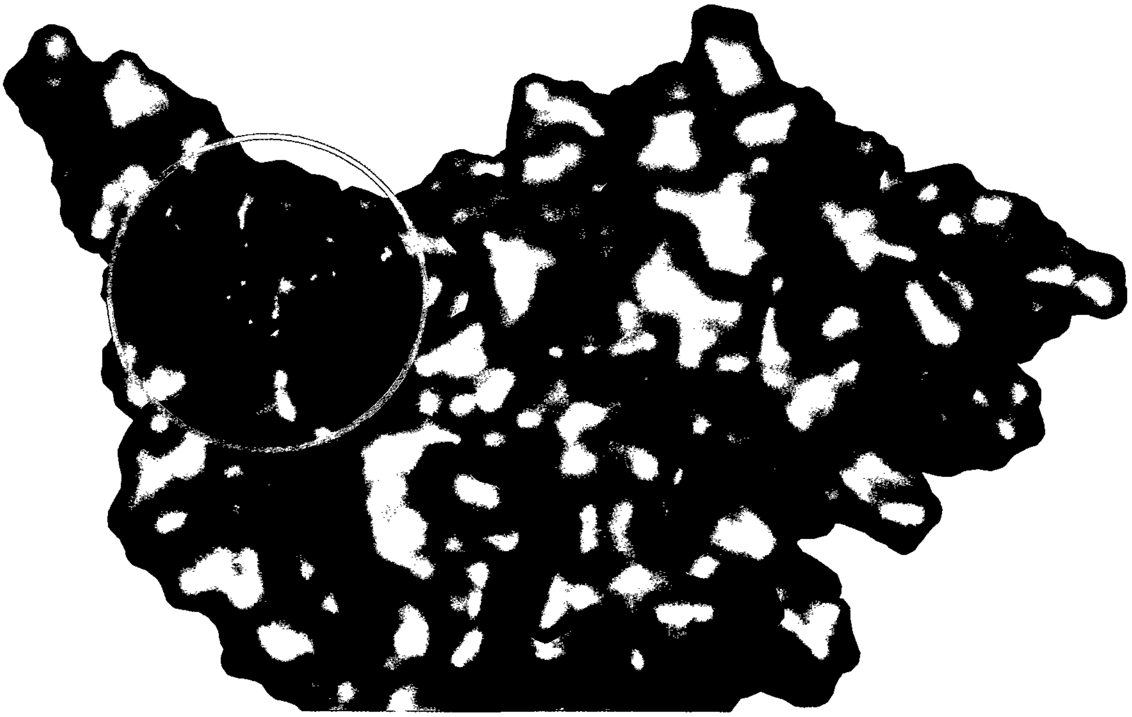


Figure 3-6: Surface of a bacterial ABC transporter protein (PDB: 1L2T).
The protein binds an ATP molecule in the pocket on the upper left corner. A site-sphere (represented by the yellow circle) was defined and the surface within this sphere (colored by the lipophilic potential) was used for the surface comparison.

usually cover only a small fraction of the total protein surface. Therefore the comparison of two proteins can be reduced in many cases to the comparison of their binding sites. This will allow the investigation of the relevant parts of the proteins' surfaces at the same resolution as low molecular weight compounds. A potential drawback of this approach is that information about the location of the functional site is needed, but that is usually available together with the 3D structure of the protein.

There are several ways to select the region of interest around a functional site. SURFCOMP applies one of the most popular strategies. A spherical region is defined such that it encompasses all amino acid residues that are known to be part of the site. Note that in the case of elongated binding pockets other choices (such as a union of overlapping spheres) would be possible. To restrict the surface comparison to the area around a site the initial association graph is built only from those critical points, which are included in the site-spheres. The rest of the process is performed as described above.

ESP Calculation. For the calculation of the electrostatic potential on the surface of a protein (see also section 2.2.1 on p. 11) it is usually not practical to use *ab initio* or semi-empirical calculations due to the large size of the systems. One viable compromise is to assign point charges derived from protein force fields. In the present investigations the charges of the AMBER force field [36] were used.

3.12. Implementation Details

The following section gives a general overview about the programming techniques, software and libraries that have been used for the various surface comparison experiments. If any experiment required different or additional tools it is described in the corresponding section of chapter 4, "Computations and Results".

3.12.1. Software

The complete surface comparison process, as described in the sections above, was implemented in the computer program SURFCOMP. The main program and all necessary libraries were written in C++ and binaries were compiled for Linux with the GNU compiler suite [53]. All matrix and vector manipulations have been coded with the RazorBack 2.0 library [8] and all graph operations were implemented using the Boost graph library [123].

The molecular surfaces were calculated by the MOLCAD module [24] in Sybyl 6.9 [2] or alternatively the molecular surface program MSMS by Michael Sanner [114;115]. All the surface properties were calculated by the MOLCAD module. For the electrostatic potential appropriate atomic point charges were either calculated at a semi-empirical level with MOPAC [40] or at the Hartree-Fock level with JAGUAR [119].

For the evaluation of the results a plug-in for the Geomview [1] software was developed that allows a real-time 3D visualization of the surface alignments, scoring and browsing of the ranking and preparation of several output formats for publishing the results. Pictures of the surfaces and surface alignment were prepared and computed using the rendering software POV-Ray [3].

3.12.2. Hardware and Computation

The actual surface comparisons were performed on a Linux cluster consisting of 22 nodes with two 2.4 GHz Intel Xeon processors and 2 GB of RAM. The generation of the surfaces and the calculation of semi-empirical atomic point charges were carried out on a four CPU SGI Origin 200 server with 4GB of RAM running under IRIX 6.5. HF calculations and the evaluation of the results were executed on a Linux workstation with two 1.0 GHz Intel Pentium III processors with 512 MB of RAM.

Depending of the size of the problem a single surface comparison usually takes from about 75 s for low molecular weight compounds up to 2 hours for the comparison of large protein active sites. The calculation of the surfaces and their properties can take from 10 up to 120 seconds except for the calculation of atomic point charges which depend heavily on the computational level (HF, semi-empiric or force-field charges).

4. Computations and Results

The primary aim of this project was the investigation of surfaces of small molecules. The method described in chapter 3 has been explicitly designed for that purpose. As a proof of concept 8 thermolysin inhibitors were investigated that were subject to an earlier surface similarity search performed by Cosgrove et. al. [37] with their SPAt program. This dataset was also used to validate the scoring algorithm against the ranking of a flexible alignment [82]. In addition another set of structures was assembled, which contains known active ligands of dihydrofolate reductase, to test the performance of different kinds of molecular surfaces. The effects of conformational flexibility were also tested with different conformations of ATP^{4-} and of a dihydrofolate reductase inhibitor.

During the project it was possible to apply the program to protein/protein and in particular to ligand binding site comparisons. With only little adjustments SURFCOMP achieved successful and illustrative alignments between the active site surfaces of different SH2-domains and phosphatases. These alignments helped elucidating important aspects in the differences and similarities between the binding sites of these proteins.

In the following sections the experimental details and results of the aforementioned experiments are presented. Unless otherwise noted the experiments were performed according to the methodologies described in the previous chapters.

4.1. Ligand Surfaces

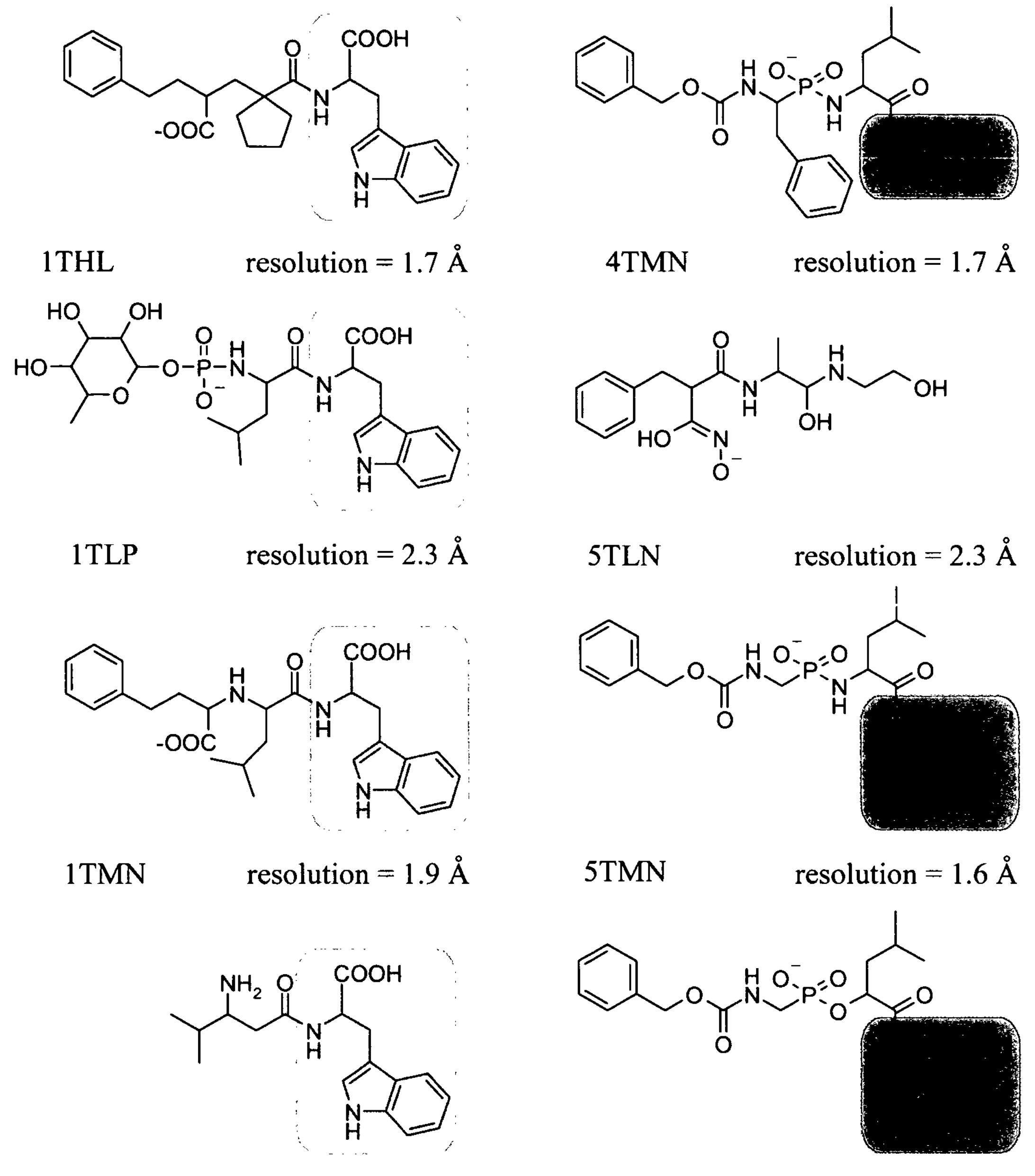
4.1.1. Preparation of the Input Structures and Experimental Design

All the molecular surfaces investigated for the experiments described in this section were calculated from 3D atomic data. The 3D structures were extracted from crystallographic data of protein/ligand complexes available in the Brookhaven Protein Data Bank (PDB) [13]. To compare the overlays generated by the present method with the experimental alignments of the different ligands in the proteins' active sites, the complexes in the PDB were superimposed by the backbone atoms of corresponding amino acids in the binding sites, which was always possible with a very small RMS deviation. The structures of the ligands were extracted and hydrogen atoms were added with Sybyl 6.9 [2].

For each structure the solvent excluded surfaces (section 2.1.4) were computed. Electrostatic potential based on semi-empirical calculations (section 3.12), lipophilic potential (section 2.2.1) and two sets of canonical curvatures together with shape type indices for a cutoff range of 1.0 and 2.0 Å (section 2.2.3) were mapped onto the molecular surfaces. For proteins the cutoff ranges were 2.0, 4.0 and 6.0 Å.

4.1.2. SURFCOMP Validation: Comparison of 8 Thermolysin Inhibitors

Thermolysin (TLN, EC-number 3.4.24.27) is a thermostable extracellular metalloendopeptidase containing four calcium ions from *Bacillus thermoproteolyticus.* [70]. The active site of the enzyme (see Figure 4-1) consists of two subsites: a zinc ion complexed by two histidine residues and one glutamic acid representing the catalytic reaction center, and a hydrophobic cleft, formed between two α-helices, that contains the selective part of the site. The crystals of thermolysin contain a lysine-valine dipeptide in this pocket that seems to be the product of the cleavage of the C-terminus of another thermolysin molecule.



1THL resolution = 1.7 Å

4TMN resolution = 1.7 Å

1TLP resolution = 2.3 Å

5TLN resolution = 2.3 Å

1TMN resolution = 1.9 Å

5TMN resolution = 1.6 Å

3TMN resolution = 1.7 Å

6TMN resolution = 1.6 Å

Chart 4-1: 2D structures of eight thermolysin inhibitors:
The structures are identified by the PDB entry name of the corresponding protein/ligand complex. The given resolution is for the complete protein/ligand complex in the X-ray data.

Several structures of TLN cocrystalized with different inhibitor compounds are available from the PDB. Cosgrove et. al. used a subset of 8 inhibitor structures to demonstrate the abilities of their molecular comparison software SPAt [37]. The same set was used to perform an exhaustive pairwise similarity search between the molecular surfaces and the results were compared with the results of the aforementioned publication to validate the program SURFCOMP.

The structures of the eight thermolysin inhibitors in Chart 4-1 were extracted from the PDB. All molecules except 3TMN and 5TLN are complexed via a negatively charged carboxyl- or phosphate-like group to the zinc ion in the active site of the protein. Thus a single negative formal charge was placed at these positions. 5TLN is also complexed to

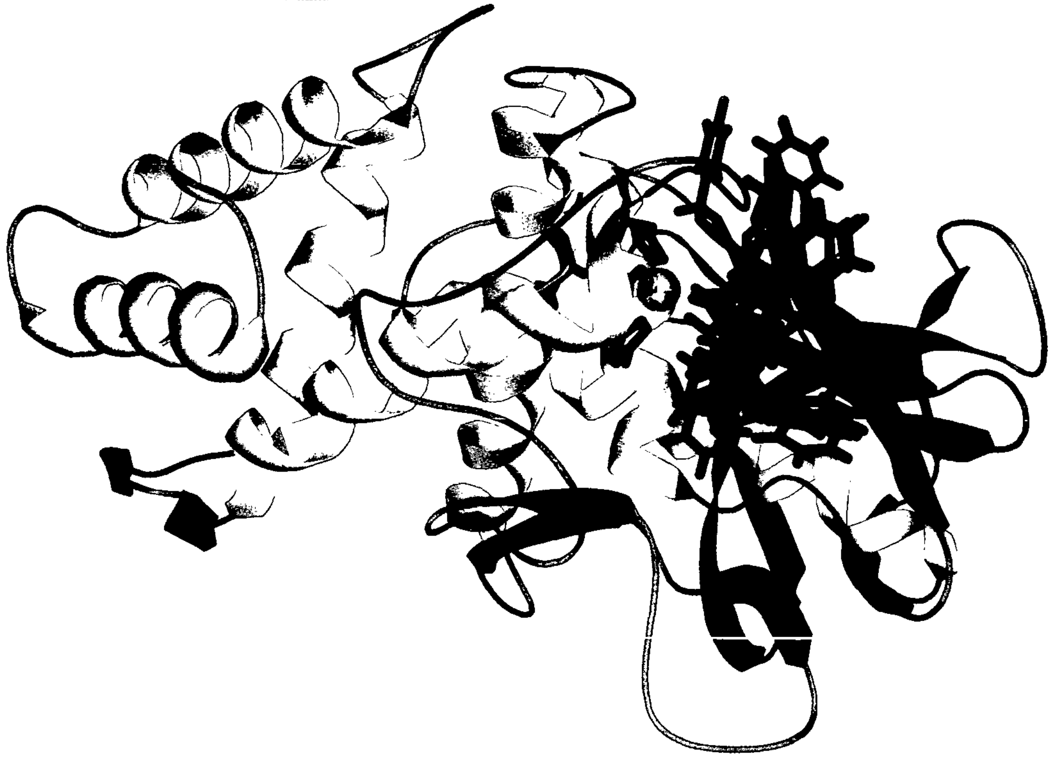


Figure 4-1: The 3D structure of thermolysin (TLN).
The 8 ligand structures of the set are superimposed in the active site. Ligands of the tryptophan class are colored in blue while the structures belonging to the valine and alanine class are shown in red. 5TLN, which does not belong to any class, is left grey. The metallic sphere represents the position of the complexed Zn ion.

the zinc ion but via a charged hydroxamic acid group. 3TMN does not show any complex binding to the ion at all and was left uncharged.

Two different experiments were performed, one with the electrostatic and one with the lipophilic potential mapped onto the molecular surfaces. The experimental details can be found in Table 4-1. Using the ESP the program could find good overlays for all structures, except for 5TLN, which is quite different in shape, especially in the most interesting region around the complex-building part. The rest of the molecules can be divided into two classes: structures with tryptophan (blue boxes in Chart 4-1) and structures with an aliphatic (alanine, leucine; red boxes in Chart 4-1) residue at the C-

filter parameter	symbol	section[a]	value	property[b]
curvature cut-off range	c_{CR}	2.2.3	2.0 Å	
neighbourhood radius	r_{CP}	3.2	2.0 Å	
fuzzy threshold	F	3.5	0.3	ESP or LP
shape threshold	R	3.6	0.6	STI
distance tolerance	T	3.7	1.0 Å	
minimum distance	δ_{min}	3.7	0.5 Å	
angular tolerance	ϕ_{tol}	3.8	15.0 °	

Table 4-1: Experimental conditions used in the thermolysin experiments.
[a] the section in the text where the filter is described
[b] the molecular surface property applied to the specific filter (ESP, electrostatic potential and LP, lipophilic potential).

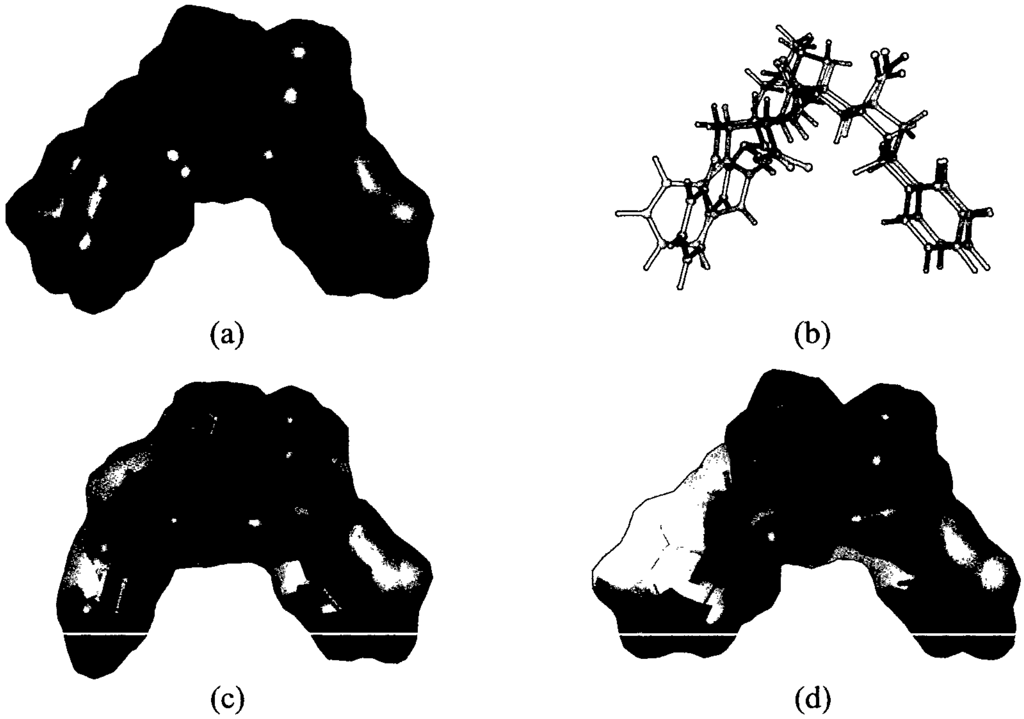
terminal end.

The tryptophan structures could be overlaid with an RMS deviation between the experimental and calculated alignment of less than 0.6 Å. The only exception is 3TMN aligned to 1TMN which shows a slightly worse RMSD of 1.0 Å mainly due to differences in their electrostatic potential and to a different angle between the indole ring and the peptide backbone. The three structures with aliphatic residues show comparable, good overlays with RMSD all below 0.6 Å. A special case is the comparison of 5TMN and 6TMN because the molecules are almost the same except for one group. Consequently their shapes and electrostatic potential are also very similar which is reflected by the small RMS deviation of 0.05 Å and the nearly one-to-one match of the surfaces.

Molecules		ESP	RMSD^a [Å]		LP	RMSD^a [Å]		Molecules		ESP	RMSD^a [Å]		LP	RMSD^a [Å]	
A	B	points^b	surf.	struct.	points^b	surf.	struct.	A	B	points^b	surf.	struct.	points^b	surf.	struct.
1THL	1TLP	580	1.95	0.40	441	1.60	1.04	1TMN	4TMN	446	1.05	1.03	417	1.44	0.80
	1TMN	711	1.77	0.40	554	1.55	0.31		5TLN	145	0.99	5.14	205	1.61	6.25
	3TMN	366	1.04	0.33	368	1.12	0.55		5TMN	464	1.21	0.93	222	0.71	0.84
	4TMN	431	1.07	1.18	349	0.98	0.95		6TMN	610	1.26	0.99	426	1.49	0.86
	5TLN	227	1.93	5.68	181	1.78	5.11	3TMN	4TMN	255	1.36	1.42	339	2.07	5.45
	5TMN	336	1.04	1.20	169	0.98	7.08		5TLN	252	1.99	2.90	116	0.58	6.91
	6TMN	439	1.00	0.63	228	0.89	0.73		5TMN	254	1.18	1.51	363	1.68	1.18
1TLP	1TMN	630	1.73	0.53	309	1.35	1.39		6TMN	180	1.26	4.28	283	1.39	0.67
	3TMN	471	1.26	0.46	424	1.52	1.20	4TMN	5TLN	383	3.52	5.83	169	1.17	6.22
	4TMN	446	2.16	1.29	188	1.51	6.01		5TMN	320	0.75	0.43	168	0.52	0.54
	5TLN	342	2.13	7.00	335	2.50	1.27		6TMN	409	0.83	0.58	312	1.90	0.49
	5TMN	454	0.93	0.63	165	0.63	1.22	5TLN	5TMN	175	1.34	2.31	176	1.44	3.37
	6TMN	409	1.12	0.59	282	0.79	1.04		6TMN	153	1.77	5.78	188	1.55	1.18
1TMN	3TMN	193	0.93	1.00	393	2.50	0.75	5TMN	6TMN	975	0.51	0.08	965	0.55	0.05

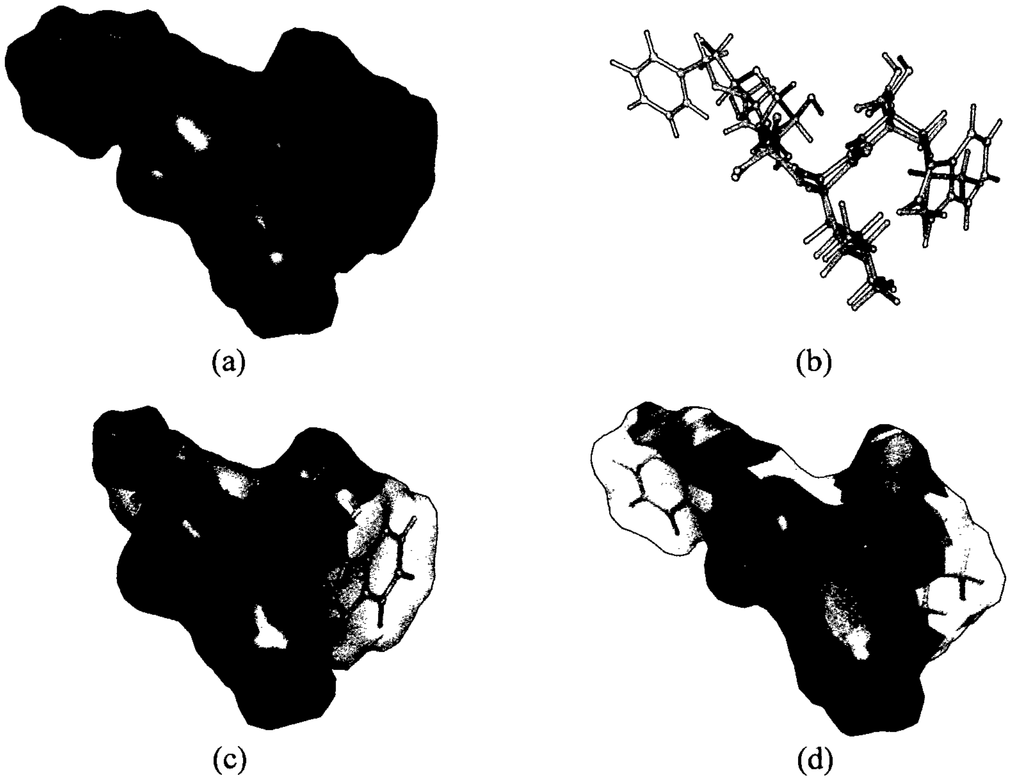
Table 4-2: Surface overlays of different thermolysin inhibitors
The surface comparisons were performed with electrostatic potential (ESP) and lipophilic potential (LP)
^{a}root mean square deviation
^{b}specifies the number of all surface points in the patches that were used to calculate the surface alignment. This number indicates the size of the similar surface region (higher number: larger region).

As expected, the overlays between the two classes were not as good as the within-class results but the general orientation and the important similar surface regions were detected correctly with RMSD values around 1.0 Å. The only exception is again 3TMN which shows rather poor alignments with the structures of the second group. This is due to the different total charge which shifts the ESP values and to the fact that 3TMN does not have the complexing group and the latter do not have the indole ring system.


(a) (b)

(c) (d)

Figure 4-3: Surface alignment of 1THL (blue) and 1TMN (red).
(a) and (b) display the alignment of the molecular surfaces and structures respectively based on the detected surface similarity. (c) and (d) show the similar surface regions of 1THL and 1TMN color coded by the electrostatic potential to illustrate their size and physicochemical similarity.

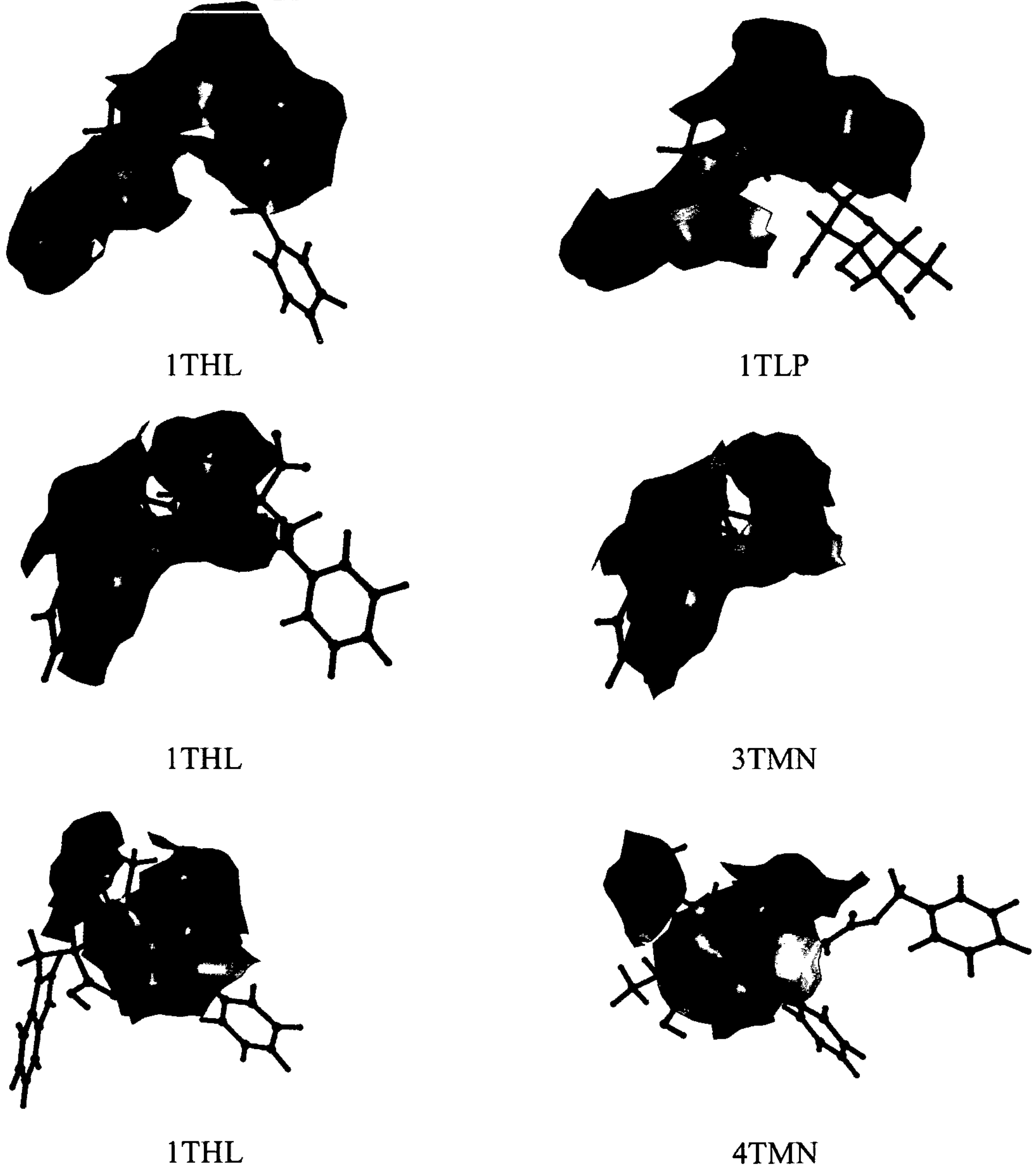

(a) (b)

(c) (d)

Figure 4-2: Surface alignment of 1TLP (blue) and 6TMN (red).
(a) and (b) display the alignment of the molecular surfaces and structures respectively based on the detected surface similarity. (c) and (d) show the similar surface regions of 1TLP and 6TMN color coded by the electrostatic potential to illustrate their size and physicochemical similarity.

The overlays found by the surface matching conducted with the lipophilic potential as the chemical filter were in general not as good as the results obtained with ESP. The main reason is that regions of the molecules that are quite close to each other in the active site, like the fructose residue of 1TLP and the phenyl ring of 1THL or 1TMN, show different lipophilicities. However the fact that the LP overlays of 3TMN on 1TMN, 5TMN and 6TMN are significantly better than the ESP overlays is due to the strong hydrophobic similarity between the alanine, tryptophan and leucine side chains. The results of both experiments are presented in Table 4-2 and example alignments are displayed in Figures 4-2 and 4-3. These results together with a description of the method have been published [65].

Besides the structure alignment based on the surfaces, the SURFCOMP program also provides a detailed picture of the surface similarities that were found between the molecules. If the results of the comparison between 1THL and the rest of the dataset are lined up (excluding 5TLN which does not show any reasonable surface similarity), one can see that the similar surface regions contain some recurring patterns (see Figure 4-4 and Figure 4-3 for 1TMN). The most common motif between them seems to be the

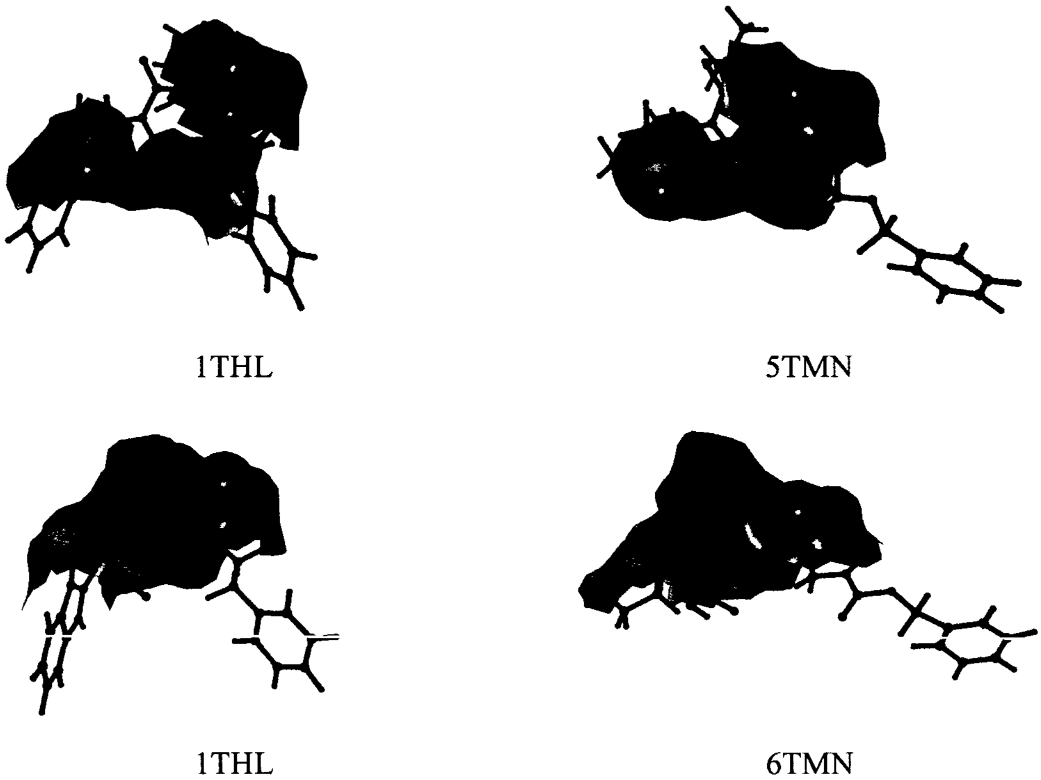


1THL 1TLP

1THL 3TMN

1THL 4TMN

Figure 4-4: continued on p. 46



1THL 5TMN

1THL 6TMN

Figure 4-4: Similar surface regions between 1THL (left) and other molecules (right).
The similar patches are color-coded by their electrostatic potential, where blue represents negative and red positive patches.

negatively charged surface region around the phosphate or carboxyl groups in the center of the molecules. The only exception is 3TMN, where that group is not present. In 1TMN, 4TMN and 6TMN the valley between this group and the C-terminal carboxylic acid is included, while in 1TMN, 1TLP and 5TMN the terminal carboxylic group itself is part of the similarity region. A strong similarity is also detected between the indole ring systems of 1THL, 1TLP, 1TMN and 3TMN where the center of that pattern is located around the nitrogen atom. Another interesting, but rather small pattern can be identified around the aliphatic sidechains upstream of the C-terminal end of the molecules. It was detected in all comparisons except for 5TMN.

The results agree with the alignments published earlier by Cosgrove et al [37] for the same dataset. The result of their SPAt program is an overlap graph, an acyclic graph that describes the best way to produce a consensus overlay between the surfaces of the dataset. In the case of the thermolysin structures, this graph consists of two connected components, which can be considered as some kind of arbitrary classification, although this is not the intention of the SPAt software. The dataset is divided into one large group containing 1THL, 1TLP, 1TMN, 3TMN, 4TMN and 5TLN and a smaller group that consists of 5TMN and 6TMN. The edges of the graph are weighted by the fraction of points of one surface that are placed within 1.0 Å of any point of the other surface by the given alignment. This evaluation of the surface similarities is different to the one used in the present experiment, because it is sensitive to differences in the size of the two compared molecules, while SURFCOMP considers only the RMSD of the similar patches. However, if the results are compared with all the data published for the SPAt calculation of the thermolysin dataset, it can be demonstrated that SURFCOMP produces comparable alignments and performs better if the size of the similar surface patches is small compared to the rest of the surfaces.

4.1.3. Ranking of Surface Alignments

If one takes a look at the number of possible alignments that are found by the SURFCOMP program in the thermolysin example (Table 4-3, below), it is obvious that a fast evaluation of the results is necessary to process a large set of surface similarity searches. In the case of the thermolysin data set the RMSD between the alignments found by the SURFCOMP program and the actual positions in the X-ray data provides a good basis for the selection of the best surface similarities. This is possible because all the structures are complexed to the same protein conformation. If, as described above, the different crystal structures are superimposed according to the backbone atoms of the thermolysin protein, the ligand molecules are brought into a natural alignment. Any superposition of two molecular structures that is based on their surface similarity can be compared to that alignment.

However, one does not always have this opportunity, especially if the 3D structures are taken from different contexts. In section 3.10 a consensus scoring scheme is described that was designed to enable a fast filtering of a native SURFCOMP result list. It produces a scoring based on the goodness of the shape fit, the size of the surface similarity and the chemical correlation, and it should be able to distinguish between promising and poor surface similarity clusters. Two different kinds of ranking are of particular interest:

(1) the identification of promising clusters within a single surface comparison and

(2) the ranking of a set of surfaces based on their similarity to a template surface.

In the first case the similarity, which reproduces the natural alignment best, should be close to or at the top of the ranking. This would guarantee that only a few clusters need to be inspected manually to find the optimal solution. The latter ranking type, also known as comparative scoring, must ensure that among the combined clusters of different experiments those surfaces are scored best that show the closest similarity to the template.

Identification of Promising Clusters. To find out, if the comparative scoring scheme is appropriate or not, the rankings produced for the similarity searches of the thermolysin data set were investigated. In Table 4-4 the ranks of the alignments which are closest to the experimental situations (I, closest cluster) compared with the ranks of the best scored clusters when ordered by the RMSD to the natural alignments (II, top cluster). The detailed results of this investigation revealed that the difference in the RMSD between the top clusters and the closest clusters were small for almost all cases where a reasonable similarity between the two surfaces exists. If no similarities could be established, the difference in the rankings and RMSD became larger. This was the case in the comparisons of 5TLN with all the members of the set, because of the totally divergent shape of its surface, and in some of the pairwise comparisons of 3TMN, particularly with 6TMN, where no satisfying similarity between the surfaces could be established.

In two cases the top cluster was the closest cluster (1TMN-3TMN, and 3TMN-4TMN). It is interesting to point out that these comparisons detected only weak or small alignments with an RMSD to the X-ray data above 1 Å. It is possible, that there are only a few acceptable alignments in such situations and the closest cluster does not face much competition from other candidate clusters. For instance the program produced only 44 different clusters when comparing 1TMN and 3TMN which is the third-lowest number among the calculations of the thermolysin dataset.

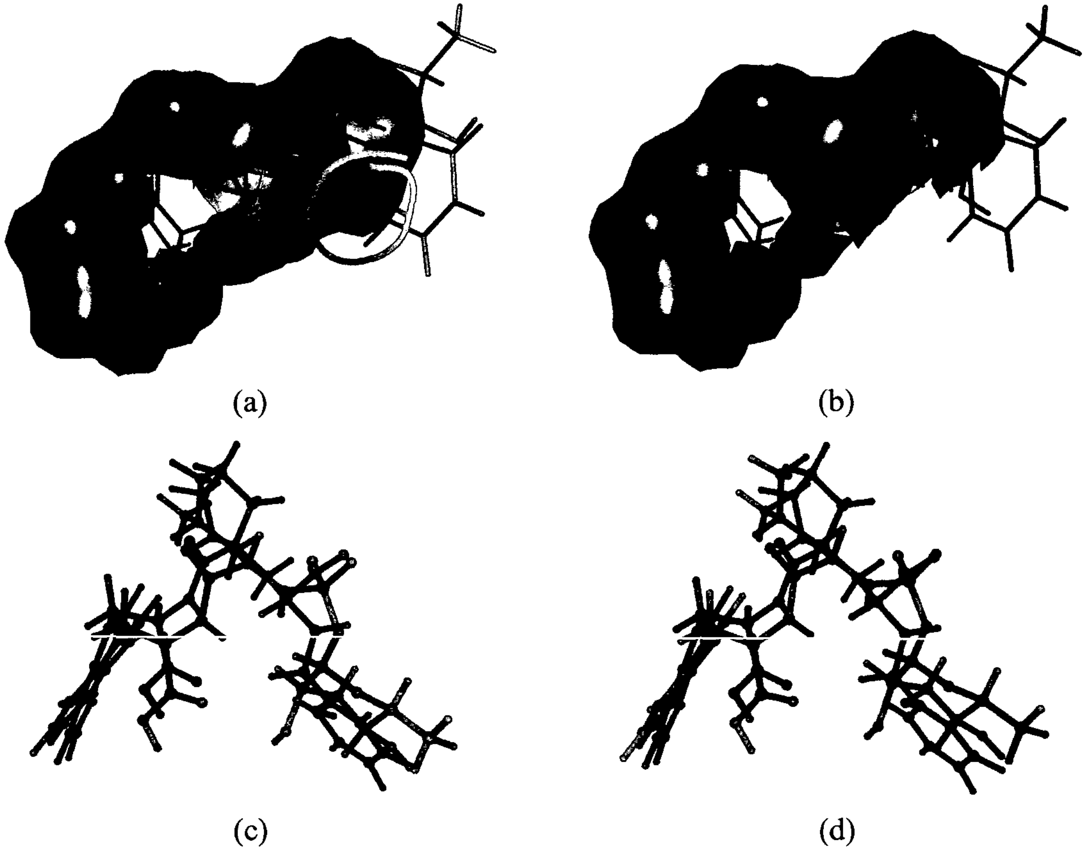

(a) (b)
(c) (d)

Figure 4-5: Comparison between the top and closest clusters for 1THL and 1TLP.
The left column shows the closest and the right the top cluster of 1THL (blue) and 1TLP (red). (a) and (b) line up the actual similar surface patches to focus on the difference between them and (c) and (d) give a snapshot of the atomic superposition viewed from the top of the molecules.

The bad ranking of the closest cluster in the comparison between 1THL and 1TLP also deserves attention. Although the difference between the top and the closest cluster is within the range of many other pairs (0.18 Å), it was ranked only at position 42 by the consensus scoring method. The main reason for that is a rather bad alignment between the corresponding surface points which is expressed by the high RMSD value of 1.94 Å. The cause of this bad alignment is a single patch pair between both molecules that could only be superimposed with a relatively large gap (see also emphasized region in Figure 4-5). The top cluster, however, does not include this patch pair and the superposition between its corresponding points is much better, although it is only the fifth best reproduction of the natural alignment (but with a very small difference). Further investigation reveals that the top cluster is a subset of the closest cluster except for the single bad matching patch pair.

Molecule				Molecule			
A	B	top-level	total	A	B	top-level	total
1THL	1TLP	19	205	1TMN	4TMN	18	132
	1TMN	15	101		5TLN	9	48
	3TMN	11	71		5TMN	14	90
	4TMN	16	131		6TMN	14	106
	5TLN	13	83	3TMN	4TMN	12	94
	5TMN	10	60		5TLN	10	64
	6TMN	10	76		5TMN	7	39
1TLP	1TMN	17	137		6TMN	9	35
	3TMN	14	92	4TMN	5TLN	18	126
	4TMN	25	201		5TMN	17	151
	5TLN	15	99		6TMN	18	160
	5TMN	16	138	5TLN	5TMN	18	52
	6TMN	22	162		6TMN	17	64
1TMN	3TMN	10	44	5TMN	6TMN	18	190

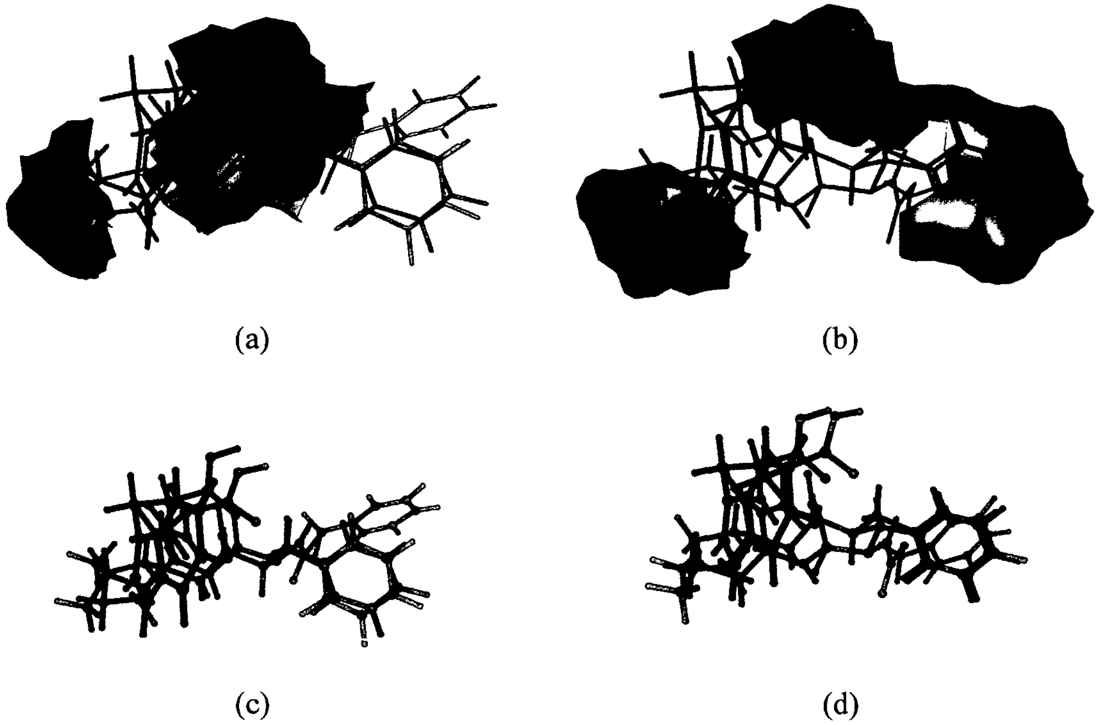
Table 4-3: The number of top level and total alignments

These clusters were found by the SURFCOMP program during the surface similarity searches in the thermolysin dataset.

Molecule					Molecule				
A	B	I	II	ΔRMSD	A	B	I	II	ΔRMSD
1THL	1TLP	42	5	0.18	1TMN	4TMN	2	16	0.76
	1TMN	5	6	0.14		5TLN	9	46	6.60
	3TMN	4	2	0.04		5TMN	8	3	0.36
	4TMN	16	22	1.21		6TMN	10	7	0.88
	5TLN	84	80	4.75	3TMN	4TMN	1	1	0.00
	5TMN	10	2	0.30		5TLN	61	24	4.29
	6TMN	8	4	0.12		5TMN	3	4	0.43
1TLP	1TMN	7	10	0.21		6TMN	21	3	1.13
	3TMN	8	5	0.15	4TMN	5TLN	46	62	3.93
	4TMN	11	6	0.52		5TMN	4	4	0.20
	5TLN	80	75	3.63		6TMN	5	3	0.20
	5TMN	7	7	0.15	5TLN	5TMN	5	19	6.16
	6TMN	4	6	0.07		6TMN	6	33	5.94
1TMN	3TMN	1	1	0.00	5TMN	6TMN	15	3	0.00

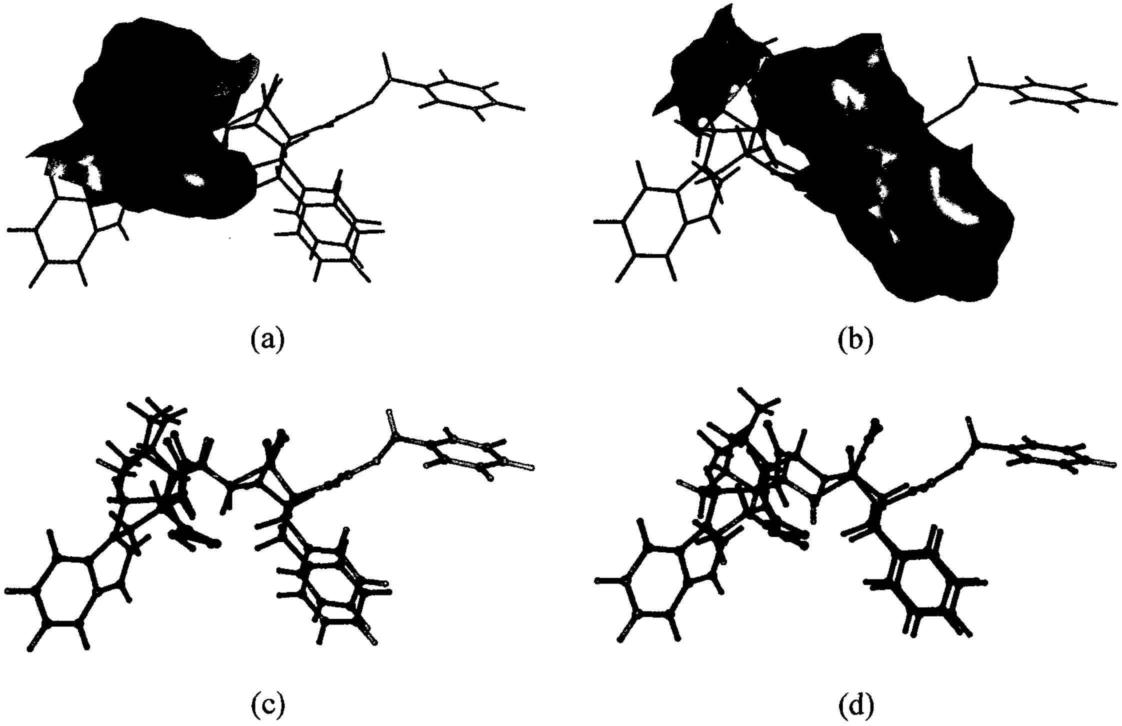
Table 4-4: Comparison of closest and top clusters.

A comparison between the ranks of the clusters that are closest to the natural alignment sorted by consensus scoring method (I) and the best scoring clusters when sorted by the RMSD to the natural alignment based on the binding site (II). In addition, the ΔRMSD between the two clusters is shown in the third column.


(a) (b)

(c) (d)

Figure 4-6: Comparison between the top and closest clusters of 1THL and 4TMN.
The left column shows the closest and the right the top cluster of 1THL (blue) and 4TMN (red). (a) and (b) line up the actual similar surface patches to focus on the difference between them and (c) and (d) give a snapshot of the atomic superposition viewed from the top of the molecules.


(a) (b)

(c) (d)

Figure 4-7: Comparison between the top and closest clusters of 1TMN and 4TMN.
The left column shows the closest and the right the top cluster of 1TMN (blue) and 4TMN (red). (a) and (b) line up the actual similar surface patches to focus on the difference between them and (c) and (d) give a snapshot of the atomic superposition viewed from the top of the molecules.

Two calculations resulted in a very large ΔRMSD between the top and the closest cluster. These were the comparisons between 1THL and 4TMN as well as 1TMN and 4TMN. All three molecules have a phenyl ring attached via a two atom bridge to the core of the structure. This resulted in a very similar surface region at that end of the molecules, which was detected by the program and used to create the atomic superpositions. Unfortunately these rings are not aligned in the active site of the proteins, which causes a displacement that is emphasized even more by the fact that the rings are placed at the perimeter of the molecules. The superpositions of the X-ray structures, the top and the closest clusters are shown in Figure 4-6 and Figure 4-7.

Finally it should be mentioned that the ΔRMSD between the top and closest clusters as well as the RMSD values between the natural alignment and the closest clusters are well below the resolution of the X-ray structures for those cases where surface similarity could be established.

Comparative Scoring. The second scoring task is the identification or ranking of the most similar surfaces compared to a template. This is similar to the evaluation of docking results, where the docked conformations of the ligand dataset are ranked to identify the most promising compounds. To accomplish this not only the alternative clusters of a single surface comparison, but the complete results of all comparisons between a set of surfaces and the given template surface are ranked by the scoring algorithm. The single surfaces of the dataset are finally sorted according to their best ranked cluster in that evaluation.

The proof of concept for that procedure can be provided by any other technique that can rank different surfaces according to their similarity to a specific template. Unfortunately, to the author's best knowledge, there is no method available that can rank molecules according to their surface similarity. Therefore the software FlexS [82] was used to validate the comparative scoring scheme of SURFCOMP, because it uses a volumetric technique to generate good flexible alignments between different molecules.

FlexS is closely related to the flexible docking program FlexX [112]. It uses various forms of possible intermolecular interactions as well as different property distributions (such as partial atomic charges or the H-bonding potential) to generate a flexible superposition between a template and a query molecule. Interaction centers and geometries or pairwise intermolecular interactions are used to evaluate the coincidence of H-bonds, salt bridges or lipophilic interactions in an alignment between the two structures. To check the similarity between certain molecular properties, Gaussian

filter parameter	symbol	section[a]	value	property[b]
curvature cut-off range	c_{CR}	2.2.3	2.0 Å	
neighbourhood radius	r_{CP}	3.2	2.0 Å	
fuzzy threshold	F	3.5	0.4	ESP
shape threshold	R	3.6	0.5	STI
distance tolerance	T	3.7	1.0 Å	
minimum distance	δ_{min}	3.7	0.5 Å	
angular tolerance	ϕ_{tol}	3.8	15.0 °	

Table 4-5: Experimental conditions used in the comparative ranking experiments.
a)the section in the text where the filter is described
b)the molecular surface property applied to the specific filter (ESP, electrostatic potential).

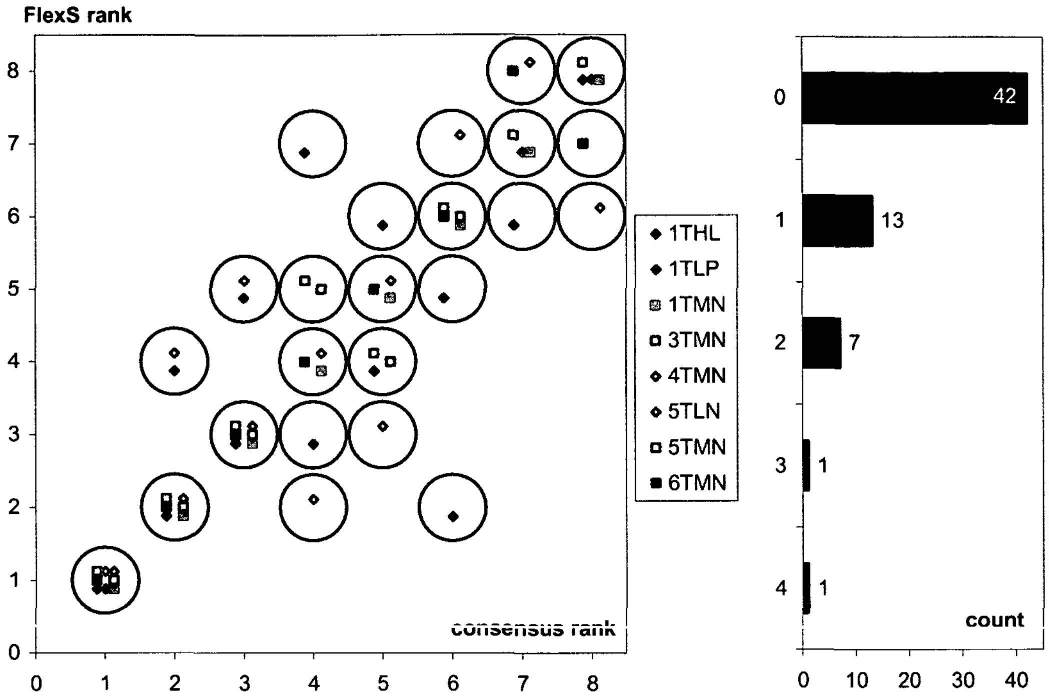


Figure 4-8: Results of the comparative ranking
(left) Mapping between the SURFCOMP consensus ranking and the FlexS ranks of all comparative ranking experiments. Each circle represents a distinct mapping between the two rankings that occurs at least once in the calculations. All correct matches appear in the diagonal of the graph. (right) A histogram of the mismatches (0 indicates a correct match).

functions that model the respective densities are used.

The experiment was designed as follows: For each structure in the thermolysin dataset, a flexible alignment with all the other structures in the set was generated. The conformations that produced the best alignment with the current template structure were taken to form the data for the surface similarity searches. Solvent excluded surfaces were generated for all structures in that set and compared to the surface of the template molecule with SURFCOMP. The resulting tables of alternative clusters were combined into one table for each template molecule and ranked by the consensus scoring approach. From this cluster scoring a ranking of the molecules of the data set was assembled based on the first occurrence of the best cluster of each molecule. The parameters for these experiments are summarized in Table 4-5.

In Figure 4-8 the results of all 8 comparative ranking experiments are summarized and the details are given in Table 4-6. Overall the agreement between the ranking based on FlexS' total score and the consensus scoring of the SURFCOMP program is very good. More than 65% of the structures were assigned the same rank by both methods and another 20% showed only a ranking difference of 1. Furthermore, many of the mismatches are still in a correct relative order. For example, the flexible superposition against 1THL ranks the molecule 3TMN at the next to last position, because it can only cover a part of the template molecule. The surface similarity ranking, however, does not take that into account, because it considers only the local similarities and does not consider the fraction of the covered template surface. Therefore 3TMN is ranked much higher by SURFCOMP because the absolute size of the similar patches is comparable to other similar molecules like 1TLP and 1TMN. It should also be mentioned that the agreement between the two scoring methods is in general better for the high and low

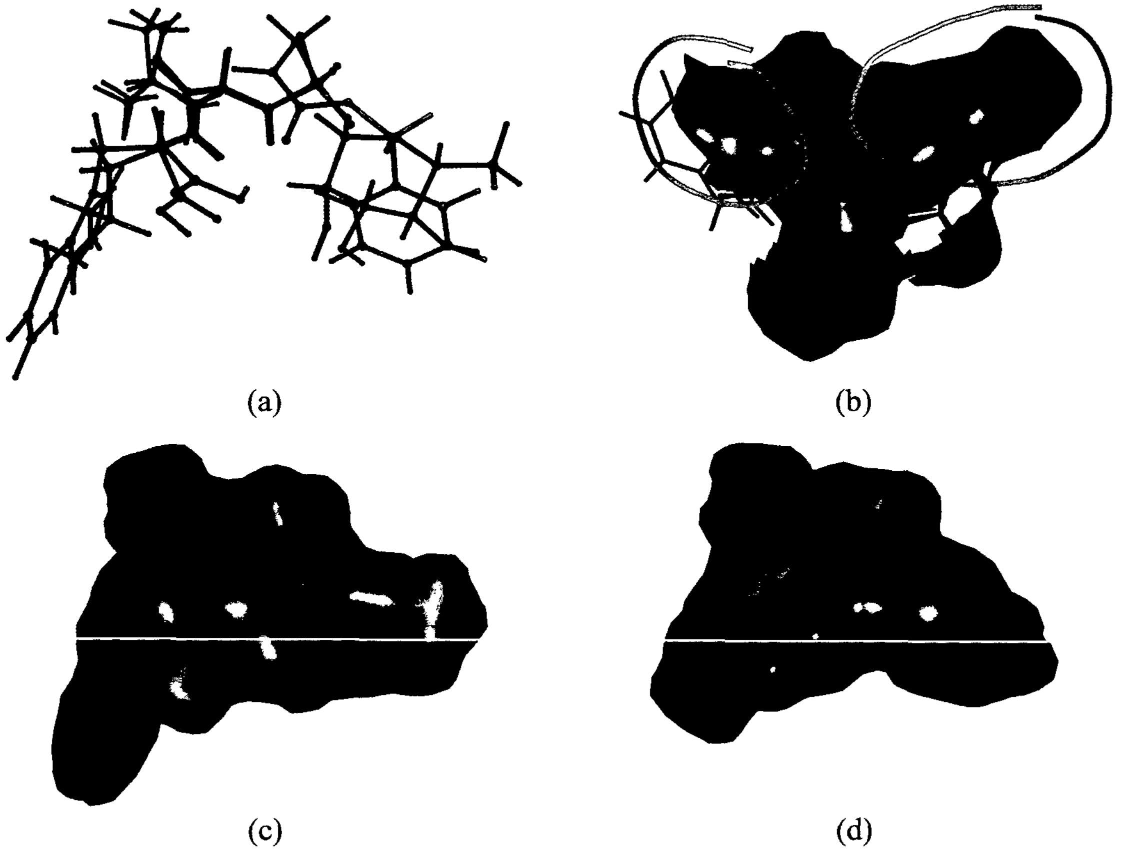


Figure 4-9: The superposition of 1TLP (blue) and 5TMN (red).
(a) The FlexS program aligns the C-terminal residues as well as the fructose and phenyl residues respectively. The SURFCOMP cluster that covers most of the surface similarities (b) has a very large surface RMSD although it represents the original superposition best. In general the two surfaces look very similar, but they have nevertheless a different ESP distribution (c and d).

ranks. While the top ranking molecules are the same for all experiments the ranks 4 and 5 are most dispersed while the exact matches increase again at the bottom of the list.

The larger differences were mainly caused by the comparative scoring experiments against 1TLP and 5TLN. As mentioned before, 5TLN does not have any significant surface similarities with any of the other molecules, which makes a reasonable ranking based on that criterion most unlikely. The situation with 1TLP is more difficult to explain, but the main reason for the bad correlation between the FlexS and SURFCOMP rankings is the fructose residue of 1TLP and the way the rest of the molecules are superimposed to that structural feature. A good example for these effects is the behavior of 5TMN in that experiment: The superposition algorithm aligned the phenyl ring of 5TMN with the fructose moiety of 1TLP and the valine side chain with the indole ring system (Figure 4-9a). These conformational changes make the surfaces of both molecules look very similar (Figure 4-9c, d). However, a surface similarity, which is in a good agreement with the superposition found by FlexS, can only be established with a high RMSD of approx. 2.7 Å between the surface patches due to the large differences between the valine and tryptophan surface and the fructose and phenyl residues (Figure 4-9b). The best ranking cluster is a subset of the closest one, where these different parts are excluded, but it is smaller and therefore ranked after the best clusters of other molecules which are sufficiently larger (e.g. 1TMN or 4TMN).

molecules		SURFCOMP consensus scoring			FlexS	
A	B	rank.	first cluster	score	rank	total score
1THL	1THL	1	1	154.67	1	-1171.90
	1TMN	2	18	313.67	2	-1129.50
	1TLP	3	749	1037.67	3	-969.10
	3TMN	4	1170	1573.67	7	-717.10
	5TMN	5	1495	2125.33	4	-928.50
	6TMN	6	1512	2163.67	5	-872.00
	4TMN	7	1593	2376.67	6	-869.40
	5TLN	8	1677	2585.33	8	-656.00
1TLP	1TLP	1	1	227.00	1	-1425.01
	1TMN	2	18	399.67	4	-1116.86
	6TMN	3	95	603.00	5	-991.06
	4TMN	4	99	605.33	3	-1146.32
	1THL	5	270	1122.67	6	-965.99
	5TMN	6	424	1614.67	2	-1263.97
	3TMN	7	678	2275.67	7	-618.78
	5TNL	8	733	2398.67	8	-605.56
1TMN	1TMN	1	1	215.67	1	-1206.58
	1TLP	2	1008	1104.00	2	-1049.43
	1THL	3	1137	1257.67	3	-1039.12
	5TMN	4	1327	1514.33	4	-1021.14
	6TMN	5	1737	2347.33	5	-964.34
	4TMN	6	1785	2442.33	6	-908.82
	3TMN	7	2061	3067.33	7	-678.71
	5TLN	8	2326	3559.33	8	-654.66
3TMN	3TMN	1	1	22.67	1	-861.91
	1TMN	2	92	131.00	2	-773.78
	1THL	3	111	184.67	3	-758.26
	5TLN	4	157	323.67	5	-603.77
	6TMN	5	162	339.00	4	-614.55
	5TMN	6	190	402.67	6	-586.36
4TMN	4TMN	1	1	135.67	1	-1425.60
	5TMN	2	307	539.00	2	-1264.84
	6TMN	3	583	6693.33	3	-1123.36
	1TLP	4	803	1830.67	4	-991.51
	1TMN	5	1013	2471.67	5	-824.99
	5TLN	6	1034	2547.67	7	-702.75
	3TMN	7	1118	2791.00	8	-557.61
	1THL	8	1140	2846.67	6	-735.21
5TLN	5TLN	1	1	93.67	1	-739.67
	6TMN	2	13	177.33	4	-632.14
	3TMN	3	24	279.33	5	-486.10

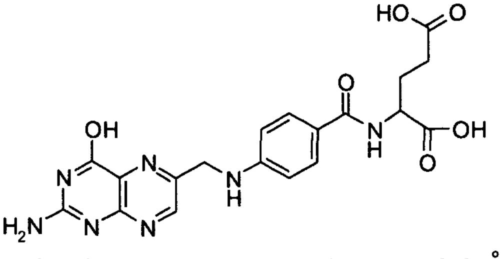
molecules		SURFCOMP consensus scoring			FlexS	
A	B	rank.	first cluster	score	rank	total score
	1TLP	4	38	359.67	2	-680.51
	1THL	5	47	383.67	3	-676.75
5TMN	5TMN	1	1	167.00	1	-1499.36
	4TMN	2	262	681.33	2	-1289.80
	6TMN	3	814	1205.33	3	-1225.49
	1TMN	4	1154	1539.33	5	-938.25
	1TLP	5	1521	1999.67	4	-1157.74
	1THL	6	2384	3658.33	6	-873.37
	5TLN	7	2394	3676.67	7	-577.19
	3TMN	8	3023	4875.33	8	-547.60
6TMN	5TMN	1	1	436.67	1	-1466.22
	6TMN	2	218	693.00	2	-1304.29
	4TMN	3	1531	1746.33	3	-1248.28
	1TLP	4	2380	2589.67	4	-1099.71
	1TMN	5	2687	3187.33	5	-883.13
	1THL	6	3016	3983.67	6	-847.09
	3TMN	7	3079	4125.67	8	-517.30
	5TLN	8	3121	4643.33	7	-568.19

Table 4-6: Comparative rankings of all molecules of the thermolysin dataset.
For the SURFCOMP ranking the comparative rank, the appearance of the first cluster of that molecule and the consensus scoring value are given. The FlexS rankings are described by the comparative rank and the total score.

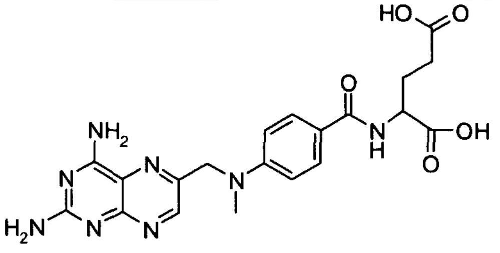
4.1.4. Evaluation of Different Surface Types: Comparing DHFR ligands

The enzyme dihydrofolate reductase (DHFR, EC 1.5.1.3) plays a key role in the folate metabolism of eukaryotic and prokaryotic cells [16]. It is responsible for the NADPH-dependent reduction of dihydrofolate to tetrahydrofolate which is required for DNA, RNA and protein synthesis. Inhibition of DHFR has been a target in drug discovery since many years and different antagonists have been developed. Methotrexate (MTX) has been successfully applied in cancer therapy and trimethoprim (TMP) is a useful drug for the treatment of various infections [121]. The triazine WR99210 is an inhibitor of malarial DHFR but shows some side effects [61]. Because of the presence of DHFR in almost any species selective dihydrofolate antagonists can be antibiotic agents as described by Li et. al. [84] for *Mycobacterium tuberculosis*. Partly because of its pharmaceutical relevance DHFR and the various folate antagonists have become a reference system for molecular modeling. Especially the DHFR/methotrexate complex is a common standard for the validation of docking algorithms [27;51;106;122;134;137;140].

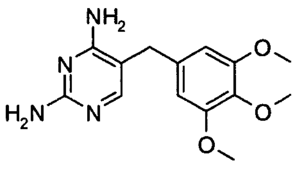
For the present investigation, a set of three folate antagonists together with dihydrofolate (Chart 4-2) was assembled. All data were taken from X-ray structures of complexes with the DHFR enzyme, which were published by Li et. al. [84] (MTX, TMP and Br-WR99210, a derivative of WR99210, henceforth referred to as WRB) and Davies et. al. [38] (folic acid, abbreviated FOL in the sequel) and involved DHFR from *Mycobacterium tuberculosis* and human cells respectively. The antagonists (MTX, TMP



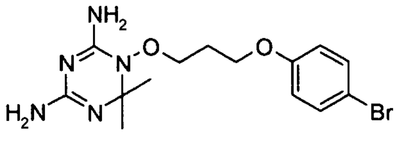
FOL (1DHF) resolution = 2.3 Å



MTX (1DF7) resolution = 1.7 Å



TMP (1DG5) resolution = 2.0 Å

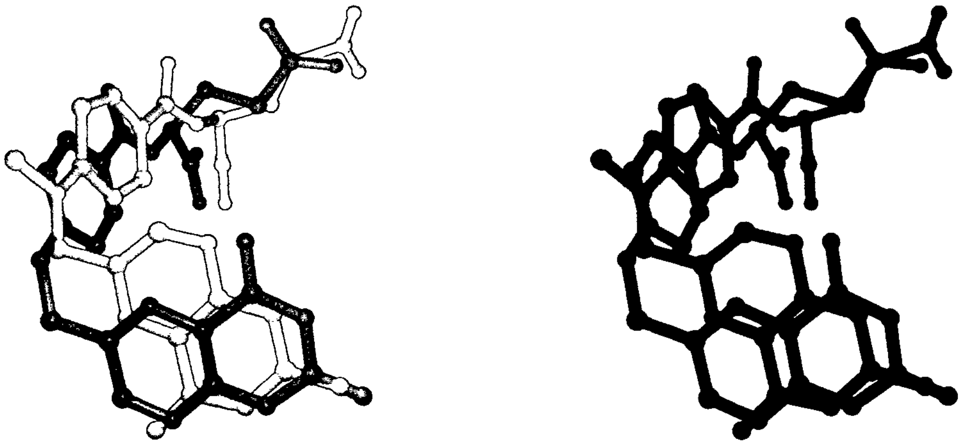


WRB (1DG7) resolution = 1.8 Å

Chart 4-2: DHFR inhibitors
2D structures of folic acid (FOL), methotrexate (MTX), trimethoprim (TOP) and Br-WR99210 (WRB). The codes in parentheses are the identifiers of the corresponding DHFR/ligand complex structures in the PDB database. The given resolution is for the complete protein/ligand complex in the X-ray data.

and WRB) were measured in a ternary complex with the enzyme and one molecule of NADPH bound to its natural binding site, which is not present in the complex of FOL with DHFR. The backbone atoms of the three complexes with the enzyme from *M. tuberculosis* (containing MTX, TMP and WRB) were aligned with an excellent RMSD of about 0.3 Å and the human protein complex with FOL could also be matched to the other protein structures with an error of about 1.0 Å. Hence the four structures could be superimposed within the binding sites of the proteins by the procedure given in section 4.1.1 on page 40.

The common feature of all four structures is a nitrogen-containing heterocycle (pyrimidine, pteridine or triazine) substituted with either one amino and one hydroxyl group or two amino groups. The remaining parts of the molecules are rather different except for MTX and FOL which have the same skeleton. When bound to the proteins the heterocycles are buried in the cleft of the active site. Several hydrogen bonds are formed



(a) (b)

Figure 4-10: Alignment of methotrexate and dihydrofolate in the pocket of DHFR.
On the left side it can be seen clearly that the two pteridine ring systems are not in perfect superposition but are rotated against each other by 60°. The right side displays the two molecules in CPK colors to show which groups are in close contact.

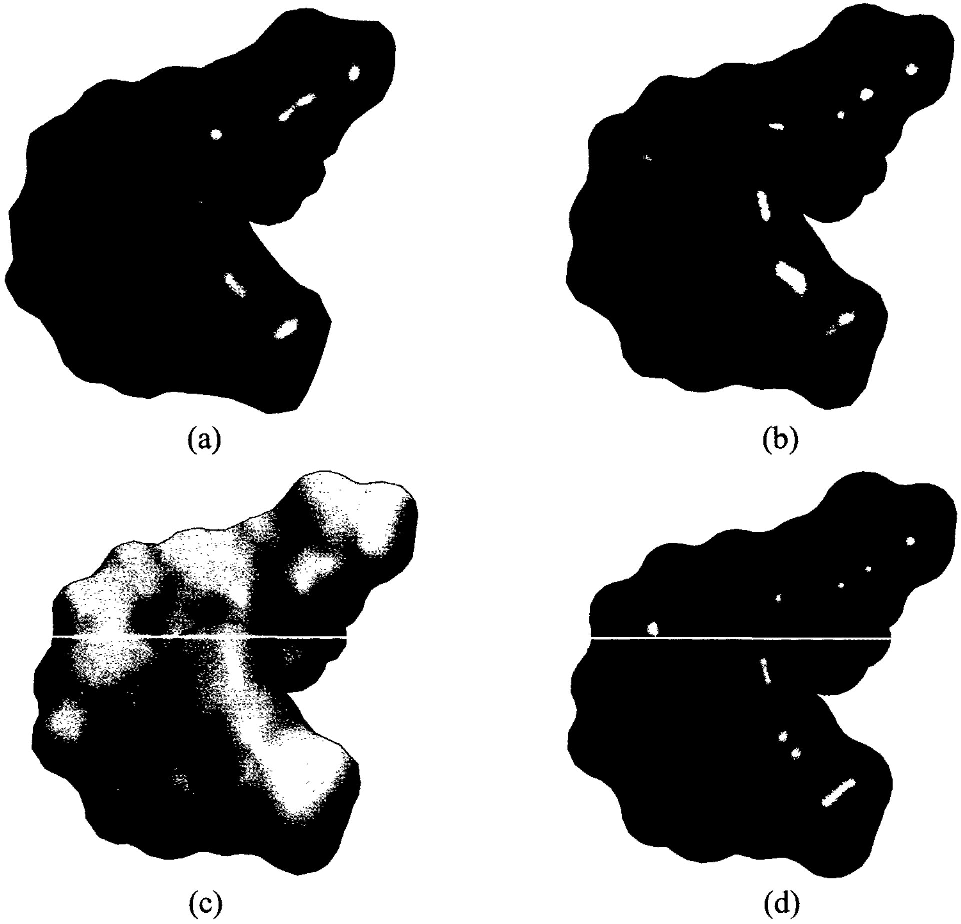


Figure 4-11: Four different molecular surfaces of the folic acid.
Fast-Connolly surfaces generated with MOLCAD with (a) 3 points per Å^2 and (b) 6 points per Å^2; molecular surface generated by Connolly's MS program with (c) 3 points per Å^2 and (d) 6 points per Å^2.

between the nitrogen atoms in the ring systems, the amino or hydroxyl groups of the ligands and different residues of the protein (especially ASP 27 ILE 5) or the NADPH molecule. The other molecules are forming different hydrophobic interactions with various amino acids of DHFR. An interesting difference in the binding modes can be observed between methotrexate and dihydrofolate. Although these two molecules have only two different functional groups (one amine is replaced by a hydroxyl group and a methyl group is added to the nitrogen that connects the pteridine with the phenyl ring), the orientation of their heterocycles is completely different. The two pteridine rings are aligned in a way that the 4-amino group of MTX is aligned with the 2-amino group of FOL. The consequence of this is that the fused rings are rotated by approximately 60 degrees against each other, while the central phenyl rings and the glutamic acids are still in a good superposition (Figure 4-10). One would not expect this constellation by comparing just the 2D molecular structures and it is a challenge for the program not to get confused by the similar looking shapes of the heterocycles.

The three different nitrogen heterocycles that form the common basis of the dataset are posing another problem to surface comparison: due to the planar character of the aromatic or conjugated systems the molecular surfaces around those parts of the compounds have only a few features that can be used as critical points in the SURFCOMP algorithm. In the case of dihydrofolate, only the amino or hydroxyl groups are responsible for a few clear peaks (see Figure 4-11a) and in the other molecules those features are even symmetric and can lead to upside-down alignments. Consequently the first preliminary experiments did not perform very well (see below). To improve the results for this dataset the features had to be enhanced, especially around the heterocycles.

One possible solution to that problem is to increase the number of points that describe the molecular surface. Increasing the point density will decrease the triangle sizes and will allow the identification of smaller features on the surface. Usually the surfaces were created with 3 points per Å^2, which corresponds to an average triangle area of 0.18 Å^2. To obtain a finer representation a set of surfaces with a point density of 6 points per Å^2 was created. Other factors that control the resolution of a surface are the placement of the surface points and the triangulation process. These parameters are usually fixed for a specific surface generation algorithm, therefore not only MOLCAD's Fast-Connolly surfaces [24] but also the output of Connolly's original MS program [32] was used, which takes longer to compute, but produces a better feature resolution.

To investigate the influence of the different surface types and resolutions on the results of the experiments four different surfaces for each molecule in the DHFR dataset were created: Fast-Connolly surfaces with (a) 3 and (b) 6 points per Å^2 and original Connolly surfaces with (c) 3 and (d) 6 points per Å^2. In Figure 4-11 the different surfaces of dihydrofolate are given as an example for the complete sets. The experimental parameters are summarized in Table 4-7. The results, which are summarized in Table 4-8, show that a significant improvement in the surface alignments as well as in the reproduction of the experimental situations can be obtained if the resolution is increased from 3 to 6 points per Å^2 or if a Connolly surface is used instead of the Fast-Connolly type.

In the initial setup, 3 points per Å^2 Fast-Connolly surfaces, FOL could only be aligned properly with MTX and WRB especially around the heterocycles, but the alignment with TMP was poorer although the amino groups at the heterocycles were aligned correctly. The detected similarities between MTX, TMP and WRB did not cover everything that could be compared and the MTX vs. WRB alignment was completely wrong because the surface of the 2-amino group of MTX was assigned to the surface of the 4-amino group of WRB and vice versa. The surface similarity between TMP and WRB was correct but could not reproduce the experimental data well because of a rather unsimilar critical point pair that was positioned over a methoxy group of TMP and the ether bridge of WRB.

The same calculations performed with high resolution Fast-Connolly surfaces lead to

filter parameter	**symbol**	**section**[a]	**value**	**property**[b]
Curvature cut-off range	c_{CR}	2.2.3	1.0 Å	
neighbourhood radius	r_{CP}	3.2	2.0 Å	
fuzzy threshold	F	3.5	0.3	ESP
shape threshold	R	3.6	0.6	STI
distance tolerance	T	3.7	1.0 Å	
Minimum distance	δ_{min}	3.7	0.5 Å	
angular tolerance	ϕ_{tol}	3.8	15.0 °	

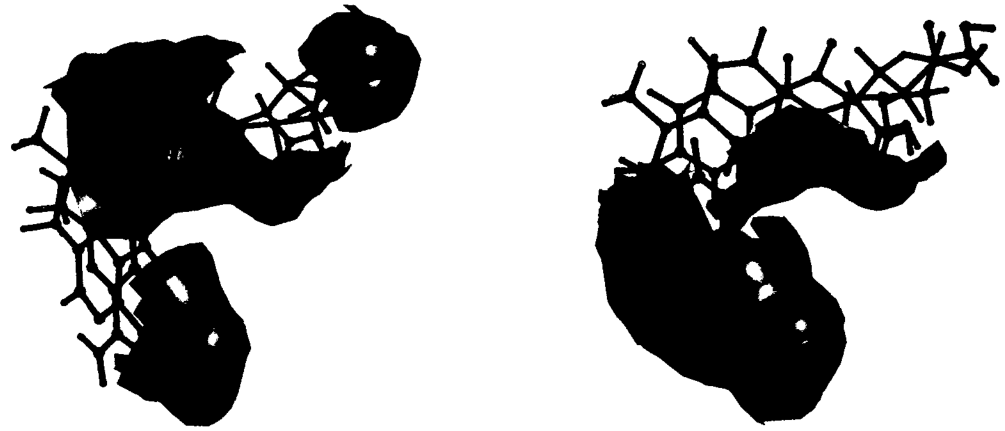
Table 4-7: Experimental conditions used in the DHFR ligand dataset experiments.
a)the section in the text where the filter is described
b)the molecular surface property applied to the specific filter (ESP, electrostatic potential).

Molecules			RMSD [Å]		Molecules			RMSD [Å]	
A	B	points	surf.	struct.	A	B	points	surf.	struct.
(a) MOLCAD surface 3 points per Å^2					*(c) Connolly surface 3 points per Å^2*				
FOL	MTX	449	1.36	1.23	FOL	MTX	372	0.85	0.73
	TMP	215	1.32	1.90		TMP	359	1.69	0.68
	WRB	257	0.99	1.26		WRB	337	0.99	1.46
MTX	TMP	216	0.64	1.63	MTX	TMP	273	1.14	0.97
	WRB	181	1.11	5.82		WRB	199	0.69	1.56
TMP	WRB	312	1.28	1.74	TMP	WRB	318	0.72	0.51
(b) MOLCAD surface 6 points per Å^2					*(d) Connolly surface 6 points per Å^2*				
FOL	MTX	890	0.76	1.61	FOL	MTX	954	0.99	1.13
	TMP	377	0.8	0.97		TMP	624	1.83	0.88
	WRB	629	1.02	1.36		WRB	721	0.84	1.58
MTX	TMP	595	1.04	0.87	MTX	TMP	581	0.53	1.37
	WRB	885	1.54	0.74		WRB	739	1.13	0.98
TMP	WRB	396	0.55	0.53	TMP	WRB	470	0.64	0.7

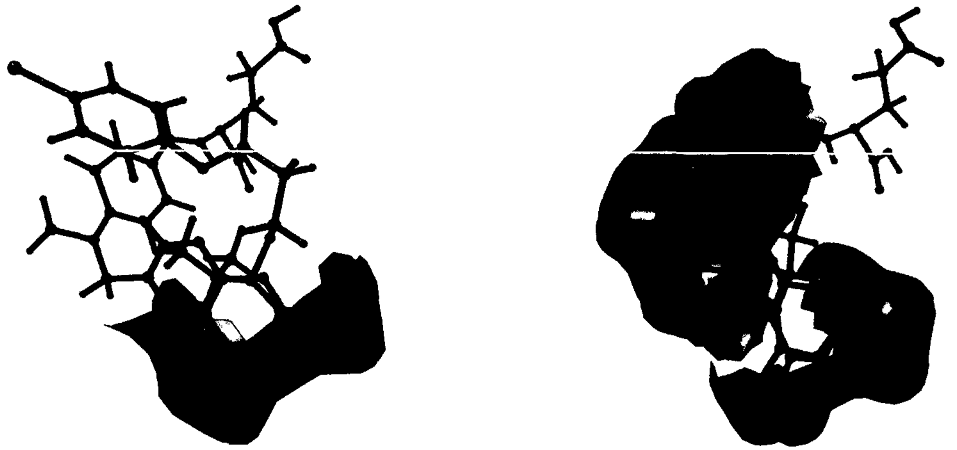
Table 4-8: Results obtained for the surface comparison with different surface types.
Under a-d are the best alignments, identified by visual inspection, for the MOLCAD and Connolly surfaces with 3 and 6 points per Å^2.

better results. For every pair except FOL and WRB, which gave the same quality, either the surface RMSD values or the displacements from the X-ray data dropped significantly. The algorithm could now find a correct alignment between MTX and WRB and the similarities between MTX and TMP were detected more completely. This usually increases the RMSD of the surface superposition, because more points are involved, but improves the fit to the experimental data. For other pairs like TMP and WRB or FOL and TMP the size of the detected surface similarities decreased because the higher resolution supported a better distinction between unsimilar pairs and therefore the representation of the X-Ray data also improved. A drawback of the increased resolution was that the calculation took up to four times longer because of the larger point sets and produced much more alternative clusters than the smaller 3 points per Å^2 surfaces of the initial setup.

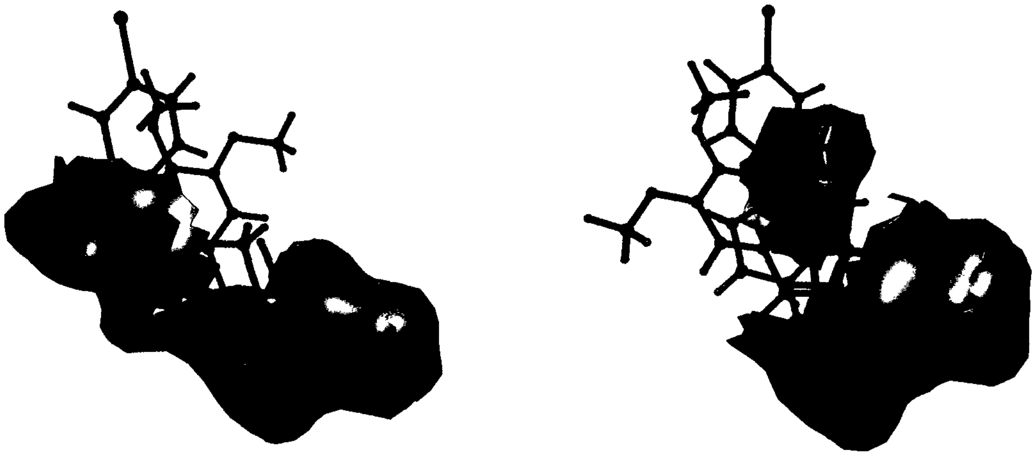
An alternative solution, which does not necessarily increase the number points, is the use of a more accurate surface type. The results obtained by the set of Connolly surfaces with 3 points per Å^2 revealed surface similarities comparable to the high resolution Fast-Connolly surfaces. In this surface type the points are placed more carefully to give a better representation of small surface features with the same number of primitives. All pairs were aligned correctly and the symmetry of the amino groups attached to the heterocycles did not cause any problems, as opposed to the case with the low resolution Fast-Connolly surfaces. The quality of the surface superposition and experimental alignment was similar to the high resolution comparisons but the patches were usually larger. Therefore some of the RMSD increased but were nevertheless of the same quality because of the increase in patch size.



FOL/MTX, Fast Connolly, 3 points per Å^2 FOL/MTX, Connolly, 3 points per Å^2



MTX/WRB, Fast Connolly, 3 points per Å^2 MTX/WRB, Fast Connolly, 6 points per Å^2



TMP/WRB, Fast Connolly, 3 points per Å^2 TMP/WRB, Connolly, 3 points per Å^2

Figure 4-12: Line ups between comparisons performed by different surface types.
The standard surface set (Fast Connolly with 3 points per Å^2) is given on the left and the improved surface sets on the right.

top: The alignment on the right side is based on a much better surface similarity that contains almost the complete area around the heterocyclic ring systems.

middle: only the alignment based on the improved surface set (right) is correct. Watch the orientation of the red structure on the left image.

bottom: presents a similar situation as in the top row; the surface alignment of the improved set is much better due to more precise surface similarities.

The last group of calculations was performed with high resolution Connolly surfaces that had a point density of 6 points per Å^2. The size of the surface similarities were slightly larger or equal to the low resolution Connolly surfaces and the computational effort was comparable to the high resolution Fast Connolly calculations. In this case the RMSD between the detected surface alignments and the fit to the X-ray data did not differ significantly from the other two improved calculations. Only MTX and TMP showed a much better surface alignment while at the same time the fit to the X-Ray data was worse than in the low resolution Fast-Connolly experiments (see Table 4-8d), because only the regions around the heterocycles were considered to be similar. The opposite was the case in the comparison of MTX and WRB. Here the higher resolution surface allowed a better identification of the similarities in the surface regions over the phenyl ring systems in both structures. Three examples that compare the results obtained by the initial setup with those of the improved surface sets are given in Figure 4-12.

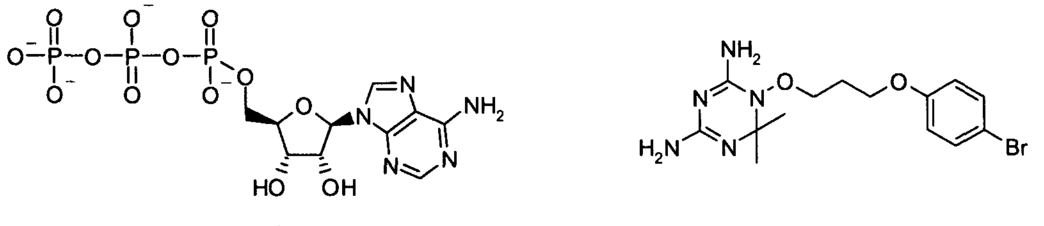
Comparing the results of group 2 and 3 (Table 4-8b and Table 4-8c) leads to the conclusion that in case of featureless surfaces an increase of the surface resolution has almost the same effects as a better placement of surface points and triangles. Increasing the point density is done easily and every surface generation algorithm provides a parameter to adjust that property. But a better point placement or a more sophisticated triangulation algorithm can usually be achieved only by a change of the generation algorithm which may not be possible in certain situations.

4.1.5. Testing different conformations

Real molecules are flexible and their actual shape can vary between many configurations that correspond to minima on the potential energy surface. The conformational flexibility of a molecule depends on several parameters including the number of rotatable bonds, the presence of rings and large groups and the environment (whether the compound is docked into an active site of a protein or is in solution). A fixed 3D structure is therefore not always a sufficient representation of a molecule but it provides all the information that is necessary to take flexibility into account. If all the atoms and bonds of a compound are known it is possible to search for new low-energy configurations on the potential energy surface using molecular or quantum mechanics.

Molecular surfaces do not provide information that is necessary to deal with flexibility. They can be seen as a view on a specific conformation that hides any information about the internal structure of the molecule. Therefore it can be difficult, if not impossible, to reproduce a surface similarity between two molecules if different conformations are used for the generation of their surfaces. To what extend the structures can vary to show still the same similarities depends on the surface comparison methodology.

The performance of the SURFCOMP program was tested on different conformations



ATP^{4-} WRB

Chart 4-3: 2D structures of adenosine triphosphate and Br-WR99210 (WBR).
The rotatable bonds considered for the conformational search are printed in red.

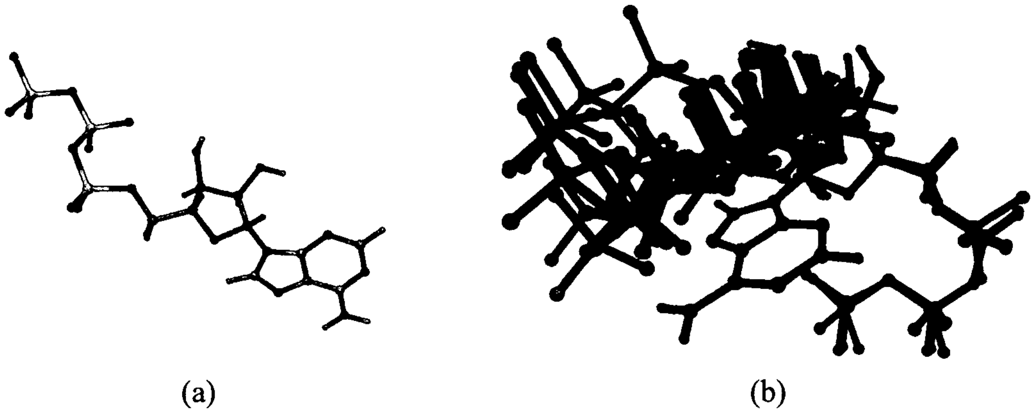

(a) (b)

Figure 4-13: Alignment of the generated conformations for ATP^{4-}.
(a) Stretched (natural) structure of ATP^{4-} when bound as a ligand to a protein and (b) alignment of the 14 different conformations as found by the random search. The conformations have been superimposed on the coordinates of the adenosine atoms. The conformations are colored according to their relative energy: blue represents the lowest energies and brown corresponds to high energy structures.

of ATP^{4-} and the DHFR antagonist Br-WR99210 (see Chart 4-3). For both molecules a set of conformations was calculated and the molecular surface of each conformation was compared with a template conformation. The detected similarities were evaluated by the result of a self-match of the template conformation, which in both cases represented an identical one-to-one association between all surface points.

ATP^{4-}. Adenosine triphosphate has usually four negative charges when bound to a protein. Therefore the three dimensional structure of ATP received four negative formal charges at the terminal oxygens of the three phosphate groups. This structure was used to generate a set of different conformations. Because of the large number of rotatable bonds a systematic search in the space of possible torsions was not possible and the random search facility of Sybyl 6.9 [2] was applied with a subset of the free bonds that includes the bond between the ribose and the adenosine, all the C-C and C-O bonds of the ribose, the connection between the ribose and the triphosphate and all the P-O bonds of the triphosphate (see Chart 4-3). The search returned 14 different conformations which were all more compact than the original conformation taken from a protein complex (see Figure 4-13). The energies of these structures varied from 28.83 kcal/mol to 32.56 kcal/mol.

With a random search the completeness of the set of conformations can be assessed by the number of times each conformation was detected by the algorithm. According to Saunders [116] the probability that a set is complete increases with the number *n* of hits for each conformation with $(1-(0.5)^n)$. Thus, if each conformation has been found five times there is a 96.9% chance that all possible conformations have been found. In the search for the ATP^{4-} molecule some clusters where only detected once in 1000 steps of the algorithm. The set is therefore not a representative sample of the available conformational space. Fortunately, for the purposes of the investigation no exhaustive list of low-energy conformers was needed, only a selection of sufficiently different shapes of the molecule.

The lowest energy conformation was taken as a template and compared with all other conformations. The comparisons were performed with Connolly surfaces at a resolution of 3 points/Å^2 and the corresponding electrostatic potentials mapped to the points (for the

Conformation	count^a	E [$^{kcal}/_{mol}$]^b	CPs^c	points	RMSD surf. [Å]^d	RMSD conf. [Å]^e
1	2	32.35	4	317	1.69	2.82
2	1	32.56	11	397	0.84	2.06
3	1	32.27	6	361	0.83	1.90
4	1	32.27	6	344	1.28	2.76
6	6	30.15	16	646	1.32	1.65
7	7	28.88	21	707	0.71	0.44
8	4	30.47	15	545	0.68	1.18
9	4	31.77	9	424	1.28	1.61
10	2	31.52	12	451	0.71	0.95
11	3	31.48	15	519	0.74	0.77
12	5	29.57	6	433	1.51	2.51
13	3	31.1	12	532	0.79	1.32
14	3	31.73	10	433	1.04	1.45

Table 4-9: Results of the surface comparison of ATP^{4-} conformations

The tests were performed with the lowest energy conformation No. 5 and all other conformations of ATP^{4-}.

$^{a)}$ number of times this conformation was detected in the random search

$^{b)}$ total energy of the conformation calculated as calculated during the random search

$^{c)}$ number of critical points that form the similar regions

$^{d)}$ RMSD between the similar surface regions and $^{e)}$ between the conformations

experimental details see Table 4-10). For each pairwise similarity search the top ranking cluster was selected by means of the consensus scoring method and the structural RMSD between the template and the test structure was evaluated as a measure of the conformational difference.

The results, summarized in Table 4-9, reveal that the size of the detected surface similarities decreases with increasing RMSD between the compared conformations. This trend is rather qualitative but it agrees with the expectations. The same trend cannot be observed between the conformational RMSD of the structures and the RMSD of the similar surface areas. For the four most similar conformations (compared to the template)

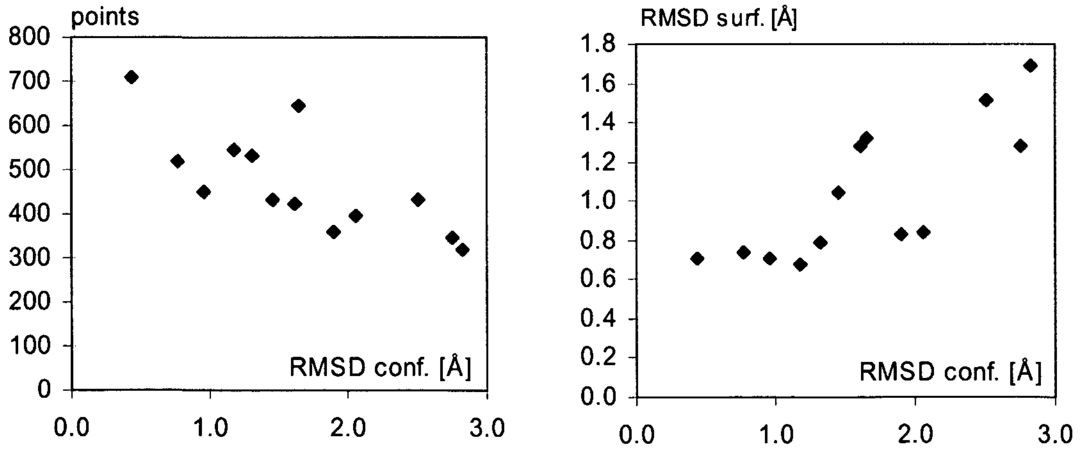
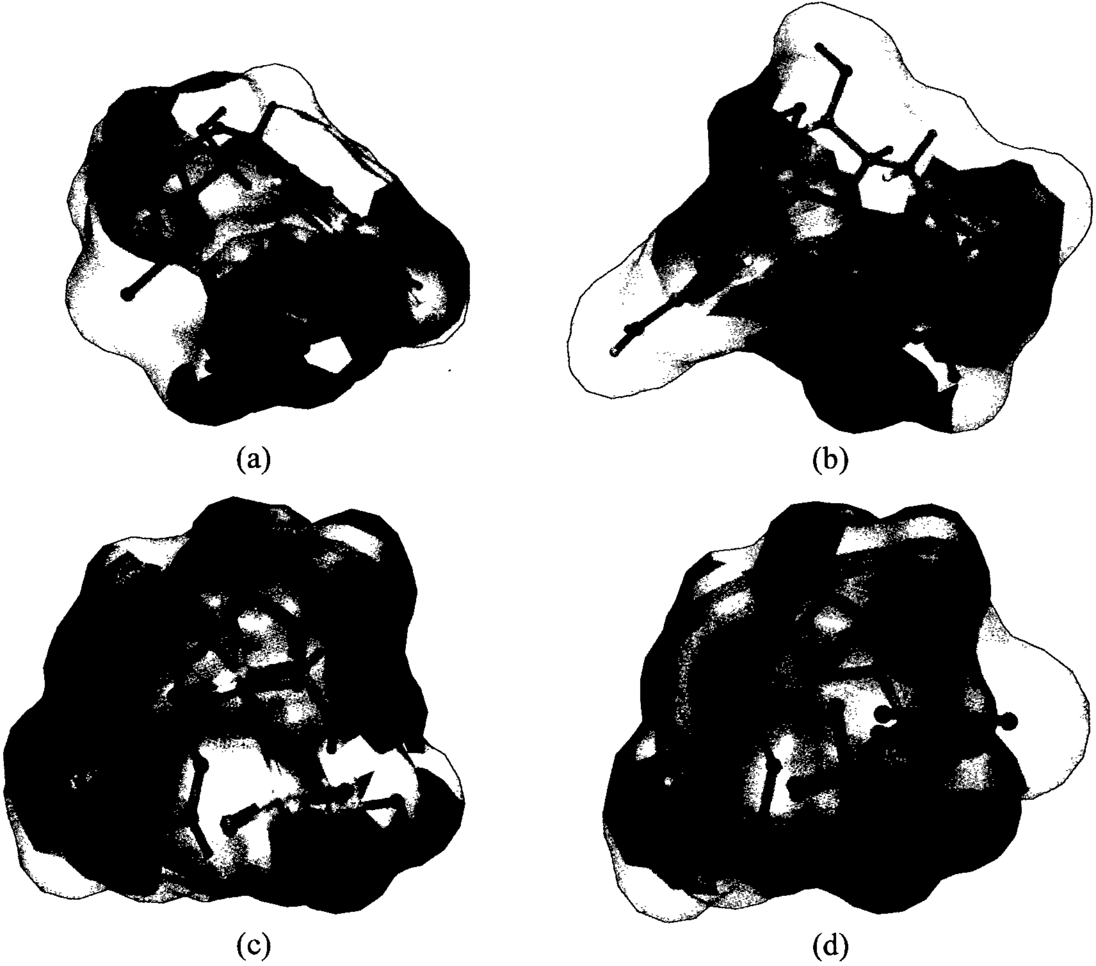


Figure 4-14: Surface similarity vs. conformational difference.

Relations between the conformational difference of the structures and (a) the size of the similar patches or (b) the goodness of the similar surface fit.

the quality of the surface fit is almost equal, and even the similar surface regions of most of the other conformations can be aligned quite well. This is because different 3D structures can nevertheless have common patches on their molecular surfaces. These regions will most probably get smaller and smaller but those parts that match can still fit very well.

Figure 4-15 shows the results of surface comparisons between two similar and two different conformations (4 and 13 in Table 4-9). It is remarkable how well parts of the molecular surfaces match each other even if the RMSD value between the corresponding structures is as large as 2.76 Å. Only the size of the patches for the less similar conformations is significantly smaller compared to the better matching structures. Furthermore, the search between the template and conformation 4 (Figure 4-15 a and b) is a good example for the case that two different structural elements can have a common molecular surface. On the other hand, the second example, comparing conformation 13 with the template, shows that although most of the structure and surface is almost identical, a single difference between the two structures, the position of the third phosphate group, prohibits the recognition of a large patch at the top of both surfaces.


(a) (b)

(c) (d)

Figure 4-15: Surface alignments of the template and two calc. conformers of ATP^{4-}.
(a) and (b) contain the template (blue) together with a bad matching conformation (red, see 4 in Table 4-9), while (c) and (d) display the alignment of the template with a well matching conformer (see 13 in Table 4-9). The surfaces are color coded by the electrostatic potential, where blue corresponds to a negative and red to a positive charge. One can see that the surfaces that match in the first example do not cover corresponding parts of the molecular structure. E.g. the positive patch on the upper left corners belongs to the ribose in (a) and to the adenosine residue in (b).

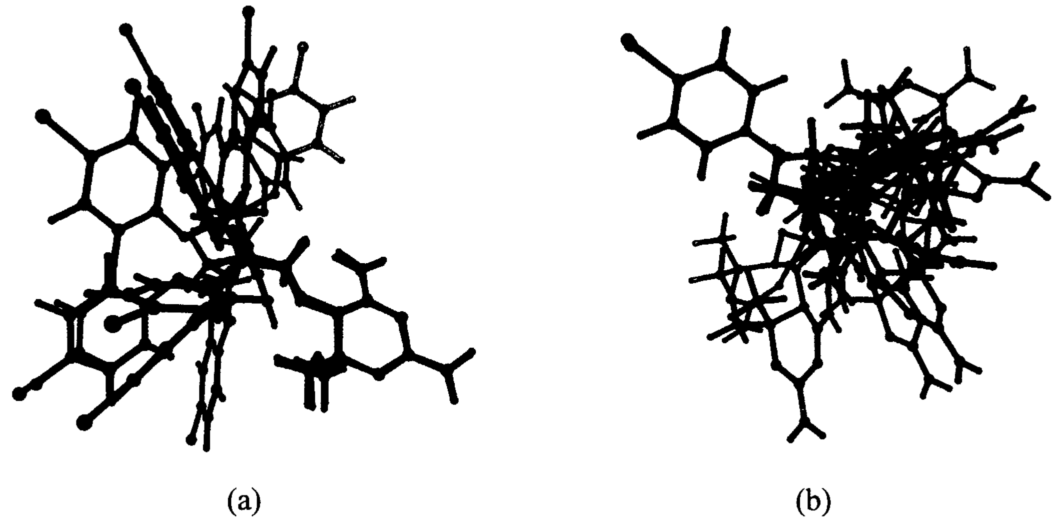


Figure 4-16: Alignment of the generated conformations for WRB.
The conformations are colored according to their relative energy: blue represents the lowest energies and brown corresponds to high energy structures. (a) Shows the relative orientation of all conformations when the structures are superimposed by the triazine rings and (b) gives the same situation for an alignment via the bromo-phenyl residues.

WRB. To investigate the actual influence of distinct conformational changes the DHFR ligand Br-WR99210 (WBR) was taken from the protein structure 1DG7. The molecule consists of two rather inflexible parts, a substituted triazine ring, which is responsible for the protein binding and a bromo-phenyl residue on the opposite end. These two parts have a very characteristic molecular surface which should be recognized easily between different conformations. The flexibility of the compound is mainly due to the ether bridge that connects the two rigid parts. Hence large changes in the 3D structure and thus in the surface can only happen in this region of the molecule. If the conformational search is focused on this area one should obtain a set of structures that will have a surface match over the rigid parts but no similarity in between. The question is to what extent these two similarities can be detected by a single surface alignment.

In this experiment a systematic search was used to generate a set of conformations. For that the two central bonds of the ether bridge were selected to rotate freely. The torsions around these bonds were changed in steps of 60 degrees which after minimization resulted in 36 different conformations having energies between 5.63 and 16.87 kcal/mol and RMSD to the original structures of 0.84 to 3.06 Å. A subset of 12 conformations is shown in Figure 4-16. From this picture one can see that the main conformational differences are the relative orientations of the triazine and bromo-phenyl residues.

With each molecule in that subset a surface comparison against the original 3D structure from the PDB structure 1DG7 was performed. The surface type and the calculation of the physicochemical properties were equal to the ATP^{4-} tests and the experimental details are given in Table 4-10. To detect how much of the surface is preserved by each conformation the results of the SURFCOMP program were searched for the clusters that included the patches around the triazine and the bromo-phenyl ring. They could be found more easily when the alignments based on the surface similarities were compared against the two different structural superpositions shown in Figure 4-16. The consensus scoring was then used to rank the clusters according to one of these RMSD differences, the size of the similar patches and the chemical correlation. This

			triazine		bromo-phenyl	
Conf.	E [$^{kcal}/_{mol}$]^a	RMSD conf. [Å]^b	points	RMSD [Å]^c	points	RMSD [Å]^c
2	8.79	1.67	452	1.19	505	0.76
9	11.36	1.15	462	0.66	378	0.56
10	6.24	1.85	485	0.63	337	0.83
13	16.87	0.84	494	0.74	395	0.69
21	10.97	2.65	436	0.68	402	0.69
23	10.75	2.21	457	0.66	392	0.89
24	6.66	1.84	483	1.17	432	1.11
25	16.27	2.79	448	0.92	377	0.97
26	10.25	2.94	330	0.53	428	0.75
27	15.65	3.06	382	0.61	373	0.62
30	10.55	2.65	392	0.79	394	0.82
32	5.63	2.65	362	0.85	437	0.79

Table 4-10: Results of the surface comparison between the conformations of WRB.
[a] total energy of the conformation calculated as calculated during the random search
[b] difference between the calculated conformation and the original structure
[c] goodness of fit of the similar surface regions.

variation to the usual scoring procedure identified in all cases the largest possible surface similarities that were centered on one of the rigid areas in the molecules.

The results show that the correlation between the size of the similar patches and the RMSD of the conformations is still similar to the relationship detected by the ATP^{4-} example and that the matches between the single rigid parts are found in every comparison. However, in almost any case – even with very small differences in the 3D structure – the two rigid areas could not be detected by a single cluster. Only if features in the ether bridge were similar they were included into the clusters that represented the conserved areas.

These results, from the ATB and WRB tests, emphasize the fact that conformational changes are a critical perturbation when two different molecules are compared. However, individual features that do not change their conformation easily are most likely detected as similar even if the total 3D structures are very dissimilar.

filter parameter	**symbol**	**section**^{a}	**value**	**property**^{b}
Curvature cut-off range	c_{CR}	2.2.3	1.0 Å	
neighbourhood radius	r_{CP}	3.2	2.0 Å	
fuzzy threshold	F	3.5	0.4	ESP
shape threshold	R	3.6	0.5	STI
distance tolerance	T	3.7	1.0 Å	
Minimum distance	δ_{min}	3.7	0.5 Å	
angular tolerance	ϕ_{tol}	3.8	15.0 °	

Table 4-11: Experimental conditions used in the conformation tests.
[a] the section in the text where the filter is described
[b] the molecular surface property applied to the specific filter (ESP, electrostatic potential).

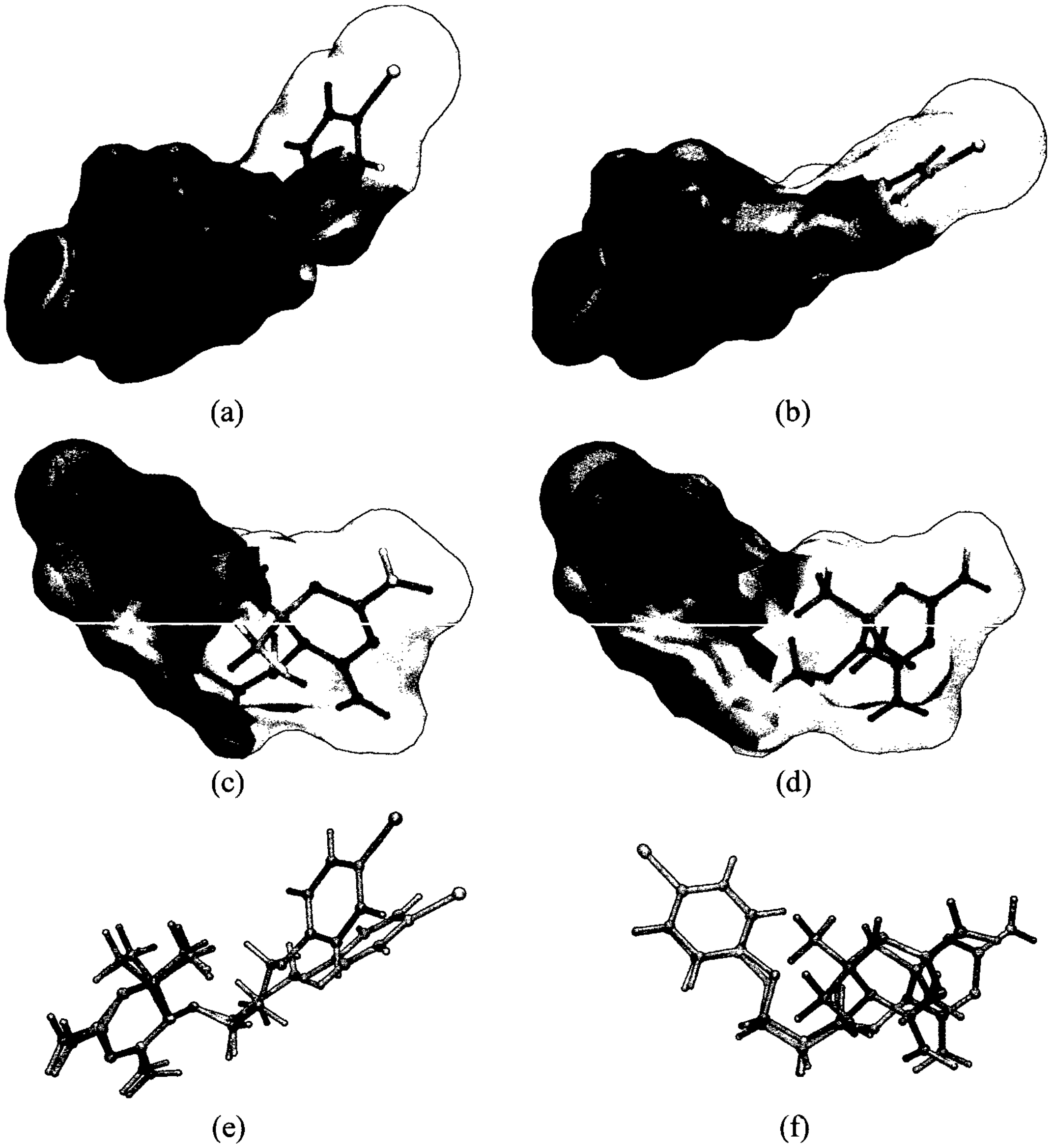
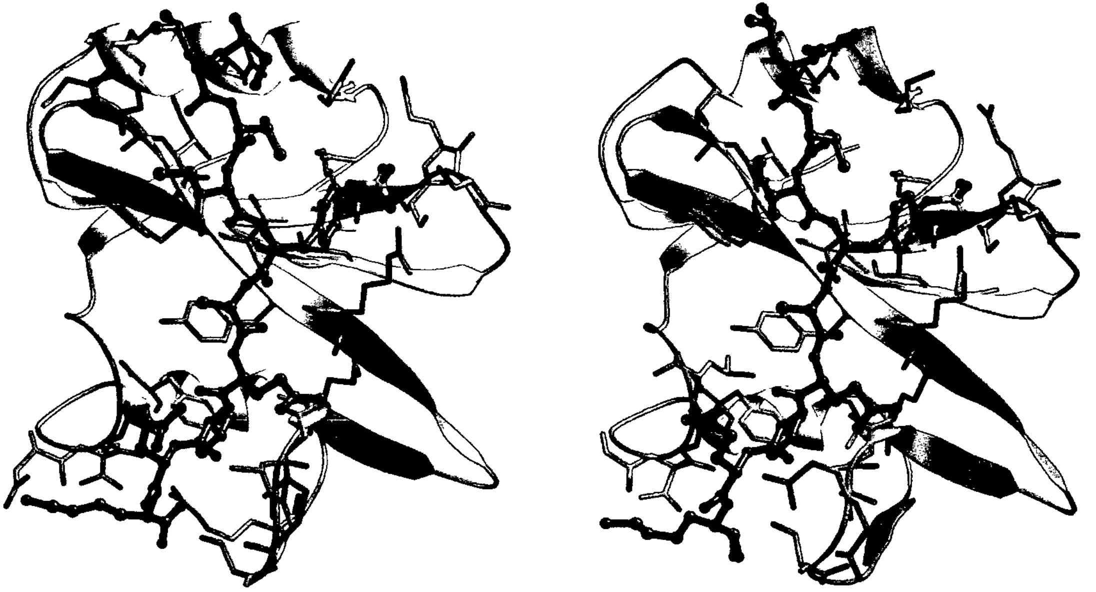


Figure 4-17: Separated surface similarities between WRB conformers.
In the top row the similar regions that matches the triazine areas are displayed (a, b), in the middle the similarity between the surfaces around the bromo-phenyl part are shown (c, d) and the bottom lines up the corresponding alignments between the two molecules based on the triazine (e) and bromo-phenyl (f) similarity.

4.2. Comparing Proteins: Surface Differences between SAP and EAT-2

SAP and EAT-2 are both representatives of SRC homology 2 (SH2) domains, which are key elements in tyrosine kinase regulation of cellular processes. The mechanism is usually triggered by the binding to peptide sequences that contain phosphorylated tyrosine residues (pTyr). SH2 domains consist of approximately 100 amino acids and can be found in a large number of proteins. Normally they can be found in higher eukaryotic cells but some evidence exists that they are also present in yeast [88]. The common fold of SH2 domains consists of a central β sheet core and a separate, small antiparallel β sheet which are flanked by two α helices, one on each side (see Figure 4-18). The phosphorylated tyrosine residue of a cognate ligand binds orthogonal to the β sheet core and residues from one side of the core and of the N-terminal α helix are forming



SAP (1D4W) resolution = 1.80 Å EAT-2 (1I3Z) resolution = 2.15 Å

Figure 4-18: Structure of the SAP-pSLAM and EAT-2 pSLAM complexes.
In both pictures the ligand peptide is displayed in bold, dark balls and sticks together with the amino acids of the protein in capped sticks that are located within 4.0 Å of a ligand atom. The PDB codes are given in the individual image captions together with the X-ray resolutions.

coordinative bonds to the ligand. The loops that connect the different structural elements can vary between the different members of the SH2 family and the affinity of a SH2 domain to a ligand peptide depends strongly on the first three amino acids that follow downstream of the pTyr [117].

When coordinated to pTyr-containing signal peptides, SH2 domains can form various protein/protein interactions with catalytic domains like tyrosine kinases or adaptor proteins like CRK or GRB2 [118]. Thereby they serve as an additional regulation mechanism in the orchestration of signal transduction that supplements the phosphorylation/dephosphorylation mediation via kinases and phosphatases. Hence, their function makes SH2 domains very interesting targets from the drug discovery point of view. Blocking SH2 domain dependent protein-protein interactions is a promising strategy for a variety of different diseases from cancer and osteoporosis to allergy and inflammatory diseases [21]. For the same reasons, selectivity between different SH2 domains is a very important factor. To avoid side effects it is absolutely necessary to target only one member of the SH2 family by an inhibitor. Therefore, the studies on the surfaces of SAP and EAT-2 concentrated on the differences of their cognate ligand binding sites.

SAP is a free SH2 domain that inhibits signal transduction events induced by a series of receptors on the surface of T lymphocytes and natural killer cells (NK). A mutation in the gene encoding SAP (*SH2D1A*) is involved in the X-linked lymphoproliferative disease (XLP), a rare immune disorder that renders the immune system unable to respond effectively to the Epstein-Barr virus [100]. SAP interacts with the consensus motif in the cytoplasmic tail of SLAM (CD150) in the phosphorylated and also in the dephosphorylated form, thereby blocking the recruitment of the SHP-2 phosphatase to that position in the receptor. Recently two groups independently discovered that the interaction of SAP with the SH3 domain of the SRC-family kinase FynT couples this kinase to SLAM [28;78].

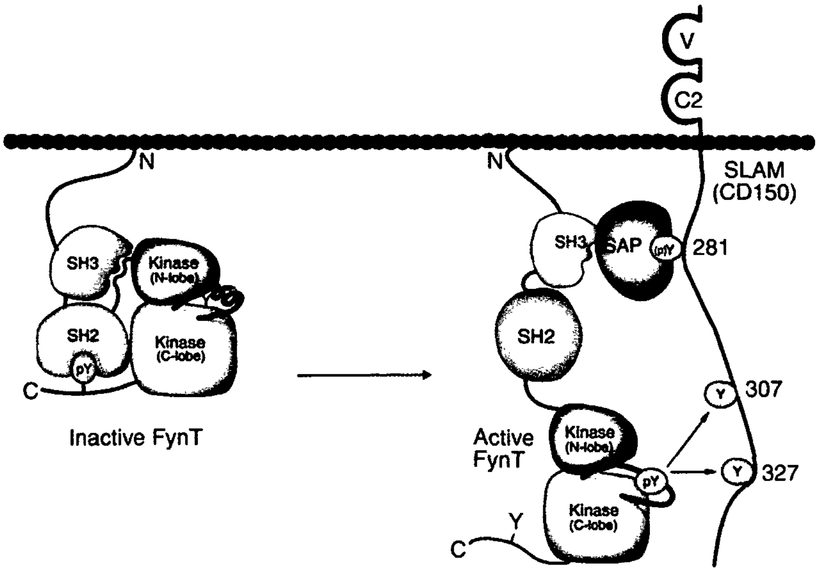


Figure 4-19: A mechanism for SLAM-induced recruitment and activation of Fyn. The inactivated form is shown on the left side and the SAP-activated form is given on the right side. (figure taken from Chan et. al. 2003 [28]).

In an experimental study, Li et. al. discovered [85] that SAP has interesting relative binding affinities to variations of the native SLAM peptide. They tested the relative dissociation constants of parts of the signaling peptide of pSLAM against the full and dephosphorylated sequence (SLAM). It was found that the N-terminal part of pSLAM is more important for the binding than the C-terminal part which is unique among the members of the SH2 family (see also Figure 4-20).

EAT-2 is a very similar SH2 domain that is expressed in macrophages and b-lymphocytes [99]. EAT-2 too can be associated to SLAM and acts as a SHP-2 blocker but no interactions with the SH3 domain of FynT are reported. Analogously to SAP, it binds to the phosphorylated cytoplasmic tail but unlike SAP it does not bind to the dephosphorylated receptor. Therefore, in contrast to SAP the binding of EAT-2 to SLAM is significantly more dependent on the tyrosine phosphorylation. This selectivity towards pTyr and the different locations of SAP and EAT-2 make the system an interesting target for a selective blocking of the SH2 signal peptide interactions.

Several protein structures for SAP and EAT-2 in the native form and in complex with the phosphorylated and dephosphorylated SLAM-tail peptide (SLTI-(p)T-AQVQK) are available. An overview is given in Table 4-12. In this study X-ray structures of both proteins in complex with the phosphorylated SLAM were used to determine the

Structure	**PDB**	**Technique**	**resolution**	**ref.**
unliganded SAP	1D1Z	X-ray	1.40 Å	[107]
SAP in complex with p-SLAM	1D4W	X-ray	1.80 Å	[107]
SAP in complex with SLAM	1D4T	X-ray	1.10 Å	[107]
SAP bound to the N-Y-C peptide	1KA7	NMR		[68]
SAP bound to the N-pY peptide	1KA6	NMR		[68]
SAP/FynSH3/SLAM ternary complex	1M27	NMR		[28]
EAT-2 in complex with p-SLAM	1I3Z	X-ray	2.15 Å	[107]

Table 4-12: Available protein structures for SAP and EAT-2.

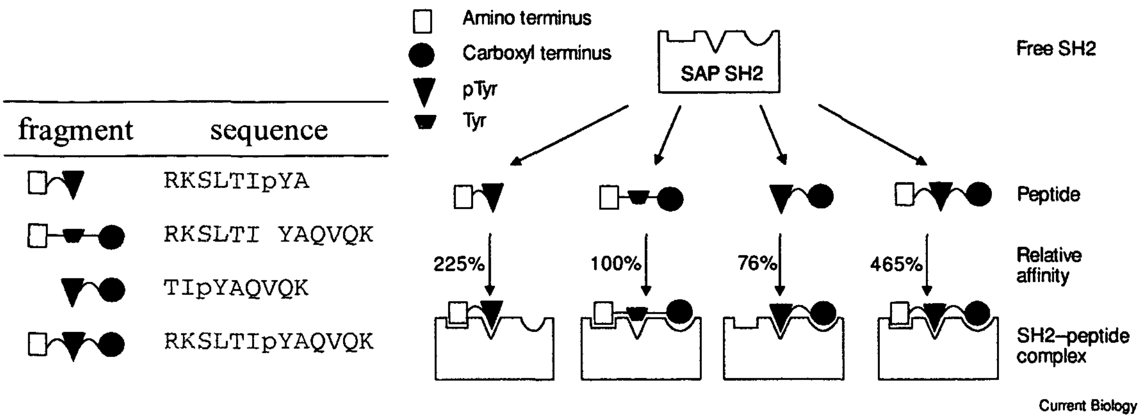


Figure 4-20: Relative binding affinities between SAP and different SLAM peptides.
The figure is taken from Li et. al. [85] and compares the relative binding affinities of different SLAM peptides with the binding sites of the SH2 domain SAP.

differences on the surface regions that are involved in the ligand binding (1D4W and 1I3Z). Sketches of both protein/ligand complexes are given in Figure 4-18 and the result of a sequence alignment is displayed in Figure 4-21.

Earlier studies revealed that the consensus sequence motive T/S-x-pY/Y-x-x-V/I is responsible for the SLAM recognition in SAP [83;85], where x represents any amino acid, and pY/Y (phospho-tyrosine or tyrosine) can be replaced by other amino acids. The three fixed residues of this motif are bound to three well formed cavities on the surface of SAP and corresponding binding pockets can be found in EAT-2. It was now of particular interest to investigate the cavities and to detect any differences in the molecular surfaces around those regions. If such differences are based on structural variations, they may highlight positions where a selective binding to SAP but not to EAT-2 could be successful. Differences due to different conformations will probably disappear if a new ligand induces a conformational change.

4.2.1. Surface Comparison

The investigation was focused on the molecular surface that was in close contact with the pSLAM peptide. Close contact was defined by selecting only those critical points on the surface which were located within 8.0 Å of the following atoms on the pSLAM peptide:

1. the carbon atom of the closer methyl group in the side chain of leucine 278 (CD1),
2. the oxygen of the hydroxyl group of threonine 279 (OG1),

```
SAP      1 MDAVAVYHGKISRETGEKLLLATGLDGSYLLRDSESVPGVYCLCVLYHGYIYTYRVSQTE
            : :·     :::·  |   |   :  ·: :|·|||| :· |    ::  :| ||
EAT2     1 MD-LPYYHGCLTKRECEALLLKGGVDGNFLIRDSESVPGALCLCVSFKKLVYSYRIFREK

SAP     61 TGSWSAETAPGVHKRYFRKIKNLISAFQKPDQGIVIPLQYPVEK
            . :  || · • .    :::: :  :   •||:·:  ·  : :
EAT2    60 HGYYRIETDAHTPRTIFPNLQELVSKYGKPGQGLVVHLSNPIMR

pSLAM  276 SLTI-pT-AQVQK

SLAM   276 SLTI-T-AQVQK
```

Figure 4-21: Sequence alignment between SAP and EAT-2.
The residues that are in close contact (6.0 Å) to the ligand peptide are displayed in blue (SAP) and red (EAT-2). A | means residue identity and **:**, • strong and weak chemical similarity.

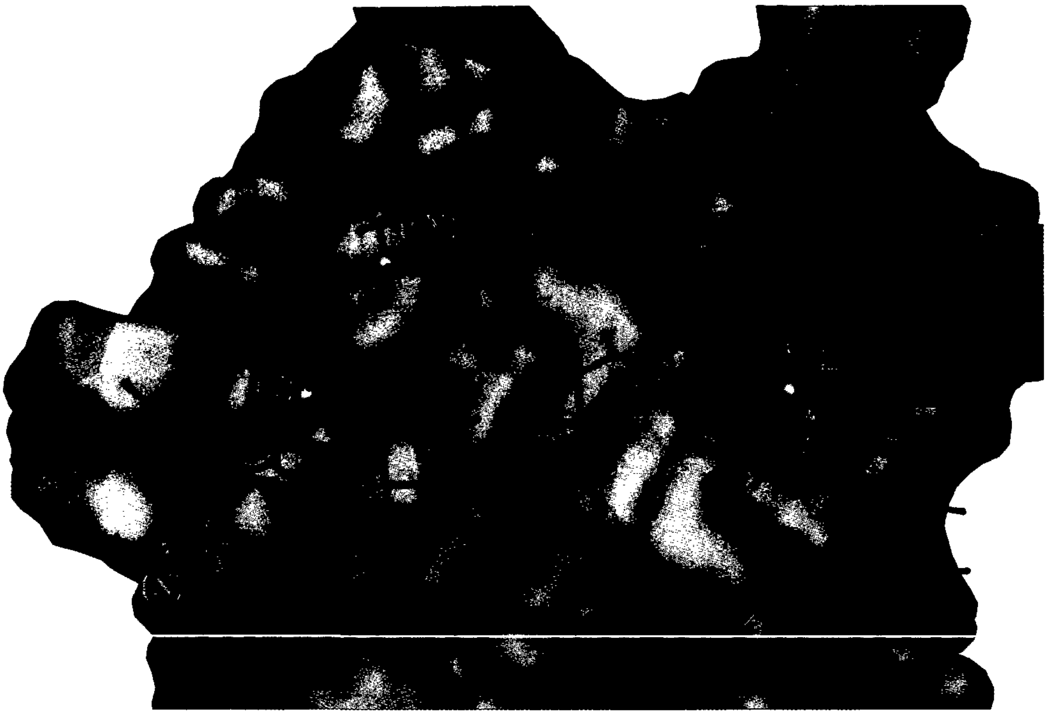


Figure 4-22: Surface regions considered in the comparison of the SAP and EAT-2.
To detect differences in the molecular surface beneath the ligand peptide, the areas on both molecules around the N-terminal residue (N), the threonine 279 (T), the phospho-tyrosine 281 (pY) and the valine 284 (V) of pSLAM were compared with each other. The yellow spheres indicate the atoms that served as central points of these regions and the blue patch defines the selected surface area.

3. the oxygen connecting the phosphate group with the sidechain of p-tyrosine 281 (OH) and
4. the β carbon in the sidechain of valine 284 (CB).

The first center represents the N-terminal part of the ligand peptide and the last three atoms are placed within the three binding cavities of the proteins that bind the fixed residues of the consensus sequence motif. Figure 4-22 shows the molecular surface of SAP with the considered regions highlighted.

A surface similarity search with SURFCOMP was performed for each of the four corresponding centers on SAP and EAT-2. To work out all the possible differences the parameters were tuned in a way to retrieve only the most significant surface similarities (Table 4-13). For the physicochemical property used in the fuzzy filtering the electrostatic potential of the protein was selected, which was calculated as described in section 3.11 (p. 37). Initially the results of each comparison highlighted only the differences in one region. To get the overall view of the complete binding area the best clusters of all four computations were combined into one picture that gives a good overview of the surface differences of SAP and EAT-2 binding to pSLAM.

Figure 4-23 and Figure 4-24 show that differences between the binding surfaces are located at the N-terminal part, at the threonine binding pocket, around the pTyr-281 location and inside of the valine-284 cavity. The central pTyr-284 binding pocket seems to be different on the upper rim and on the left side where it flanks the threonine cavity. The latter difference is mainly due to a single surface feature that corresponds to the side chain of Lys-12 in EAT-2 and the guanidine group of Arg-13 in SAP which do not show any strong interaction with the residues of the ligand. The other difference in that part, covering the binding pocket from the upper left corner is caused by different conformations of the sidechains of the glutamic acids 34 (EAT-2) and 35 (SAP). It is unlikely that these differences can serve as a starting point for a selective SAP/SLAM blocking.

Of more interest are the differences in the threonine binding pocket and the valine cavity because they cover two of the three structural motifs that seem to be responsible for the recognition of the ligand. The threonine cavity in EAT-2 is wider than but not as deep as the corresponding feature on the SAP surface. Furthermore the entrance to the cavity from the right (in Figure 4-23 and Figure 4-24) is steeper in SAP than in EAT-2. The situation around the valine pockets is even more interesting, because the differences there are larger and more complex. The finger that encloses the cavity from above the surface is much more negatively charged in EAT-2 than in SAP and the shape of that region is also quite divergent. The most important differences are found at the bottom of the pocket. There SAP has two little extra cavities that are separated by a small ridge. On EAT-2 the bottom of the valine pocket is rather flat and has no pronounced hole or ridge.

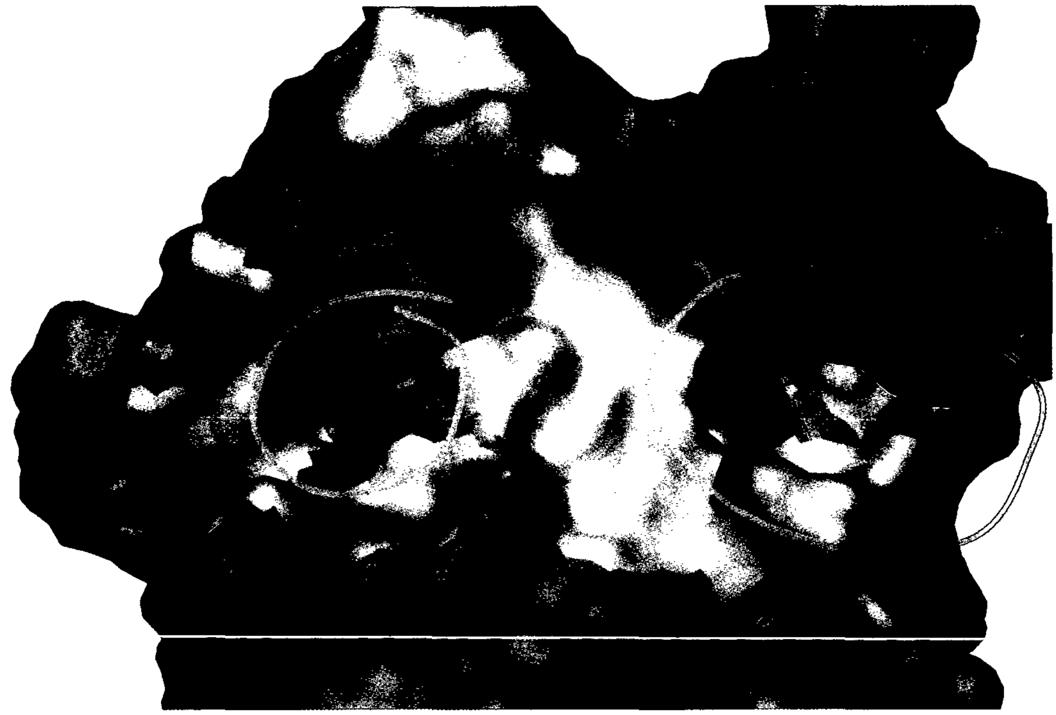
It is noteworthy that the corresponding surface patches of the proteins, which are in contact with variable parts of the consensus sequence motif T/S-x-pY/Y-x-x-V/I, are highly conserved. Neither the region beneath the Ile-280 nor the patch close to Ala-282 and Glu-282 show any significant differences, although they do not have a lot of features. These findings support the consensus motif from the perspective of the surfaces, because a flat and featureless region does not provide many anchor points which are necessary for discrimination.

4.2.2. Structural Investigations

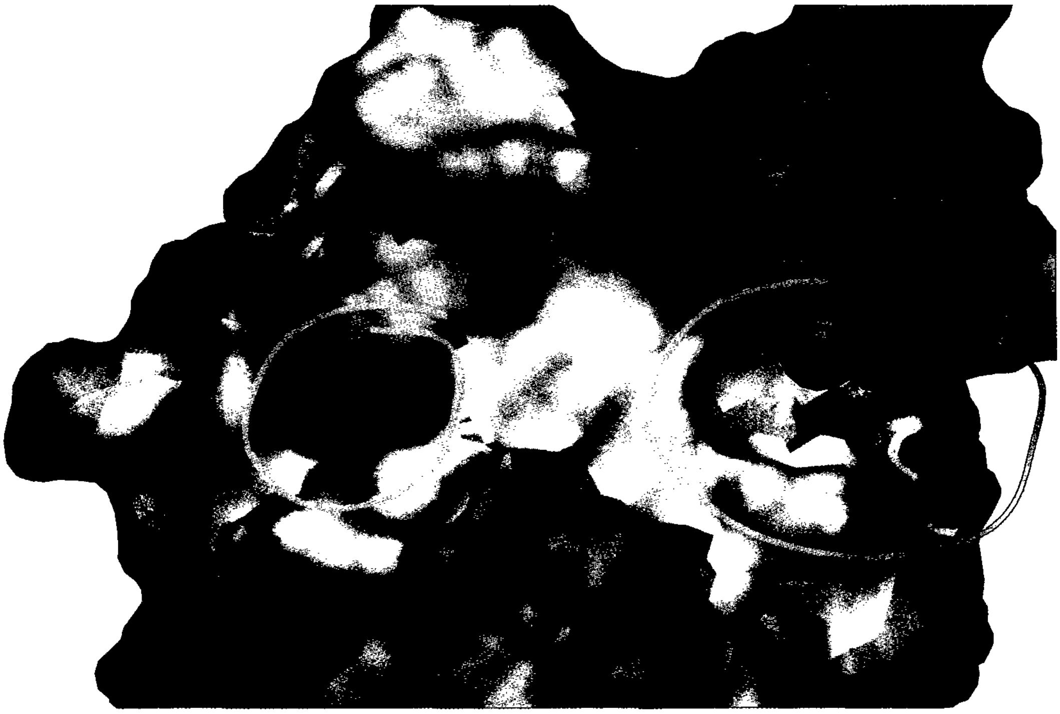
To evaluate the potential of the differences to serve as starting points for SAP/EAT-2 selectivity the structural configurations that lead to the dissimilarities in the binding surfaces have to be examined. Therefore the clusters that represented the best picture of the surface differences at each of the four sites were exported into the molecular modeling package SYBYL 6.9 [2] together with the corresponding structural and surface data. In the molecular viewer the residues that are responsible for the differences in that area could be identified easily and the surface was regenerated only for those amino acids to focus the eye of the observer on the relevant parts.

filter parameter	symbol	section[a]	value	property[b]
curvature cut-off range	c_{CR}	2.2.3	2.0 Å	
neighbourhood radius	r_{CP}	3.2	2.0 Å	
fuzzy threshold	F	3.5	0.3	ESP
shape threshold	R	3.6	0.6	STI
distance tolerance	T	3.7	1.0 Å	
minimum distance	δ_{min}	3.7	0.5 Å	
angular tolerance	ϕ_{tol}	3.8	15.0 °	

Table 4-13: Experimental conditions used in the SAP/EAT-2 comparisons.
a)the section in the text where the filter is described
b)the molecular surface property applied to the specific filter (ESP, electrostatic potential).

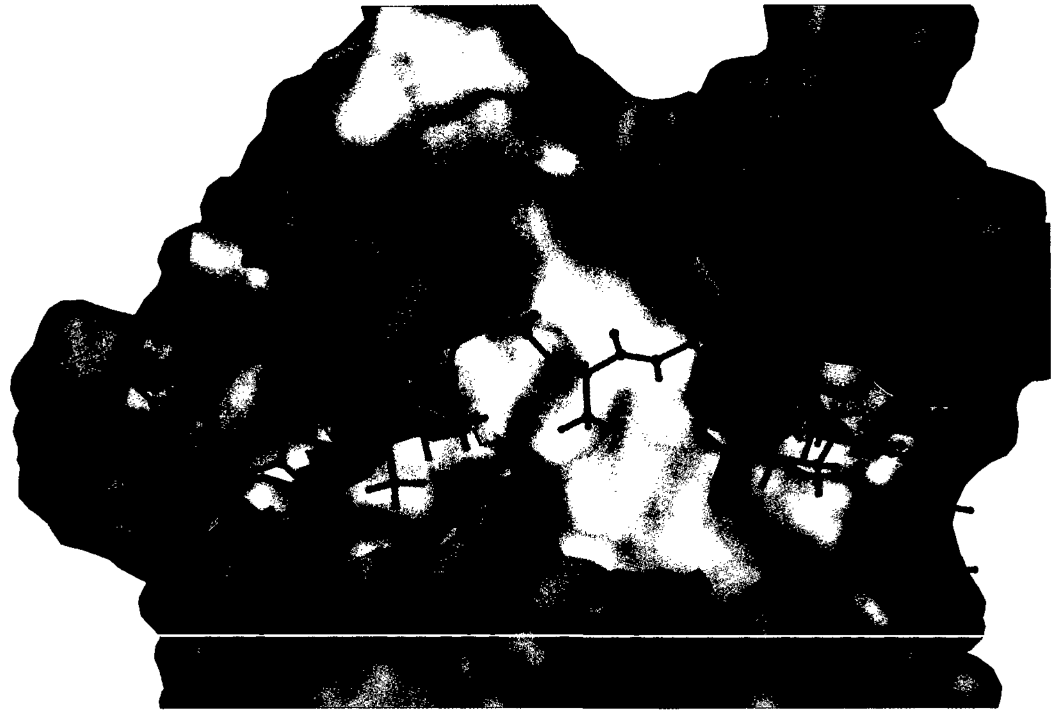


SAP

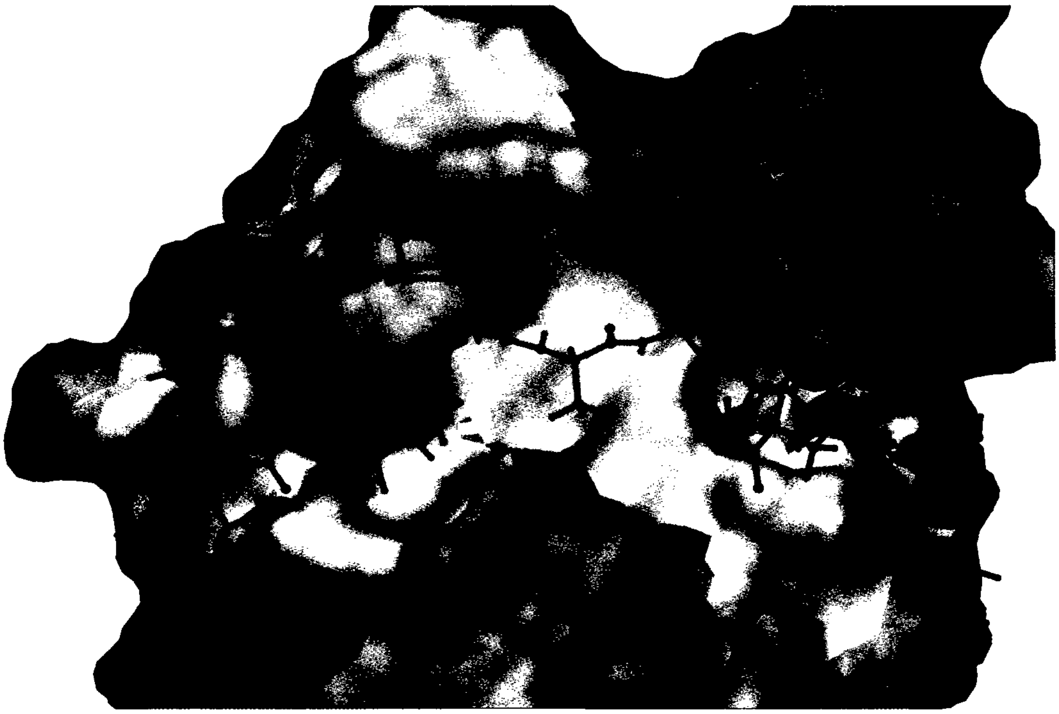


EAT-2

Figure 4-23: Surface differences between SAP (above) and EAT-2 (below).
The figure shows the differences in the surface areas that are involved in the pSLAM binding in intensive colors. The similar surface is highlighted with less intensive colors while the surface areas that were not compared are displayed in gray. The colors are encoding the electrostatic potential (ESP) of the surfaces, where blue indicates negative and red positive areas. The yellow circles indicate the dissimilarities that have been investigated in more detail on a structural level (see text).



SAP



EAT-2

Figure 4-24: Surface differences between SAP and EAT-2 with structures.
The figure shows the surfaces in the same way as Figure 4-23 but in combination with the structure of the pSLAM peptide.

The residues that form the threonine cavity in SAP and EAT-2 are very similar and the relative conformations of the residues in each pocket are also highly conserved (Figure 4-25). But the surfaces are nevertheless divergent at several points which are related to the differences in the amino acid sequence. As mentioned above a significant dissimilarity is caused by the patches that are placed around Arg-13 in SAP and Lys-12 in EAT-2. The most important difference, however, is located right at the center of the cavities where a glycine residue in SAP (Gly-16) is exchanged against a cysteine residue in EAT-2 (Cys-15). The missing side chain causes the pocket of SAP to extend deeper into the protein than in EAT-2, where the side chain of the cysteine is blocking the way. In the crystal structure of SAP the larger cavity is occupied by two water molecules which seem to be tightly bound to the protein as judged by their low B-factors of 15.25 Å^2 for the inner and 17.89 Å^2 for the outer water respectively. In EAT-2 the corresponding pocket holds only one molecule of water which is much more mobile (B-factor of 39.07 Å^2).

Similarly to the situation of the threonine cavity, the pocket that binds the Val-284 residue of the pSLAM ligand consists of some conserved and some divergent residues both in SAP and EAT-2. In contrast to the threonine cavity, the shapes of these valine cavities differ not only at the center but also at the peripheral sections. However the most

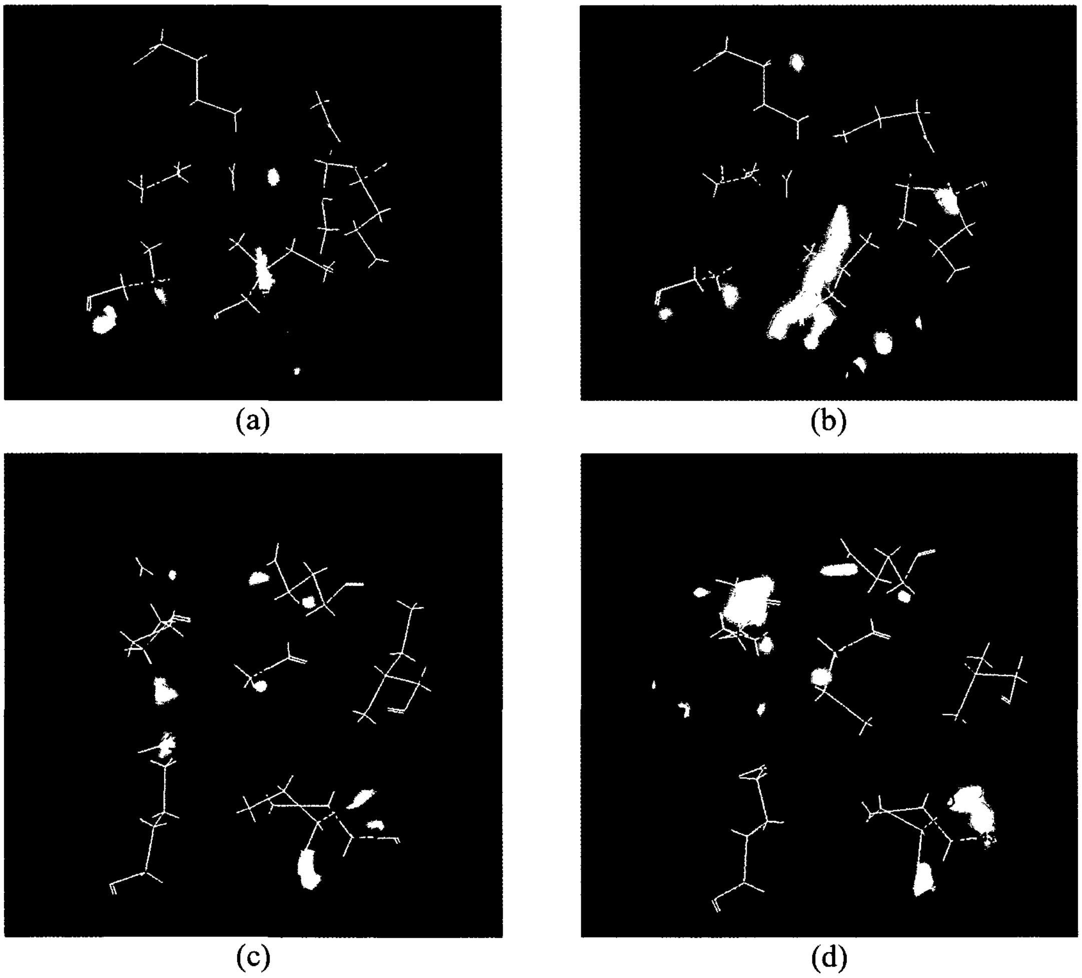
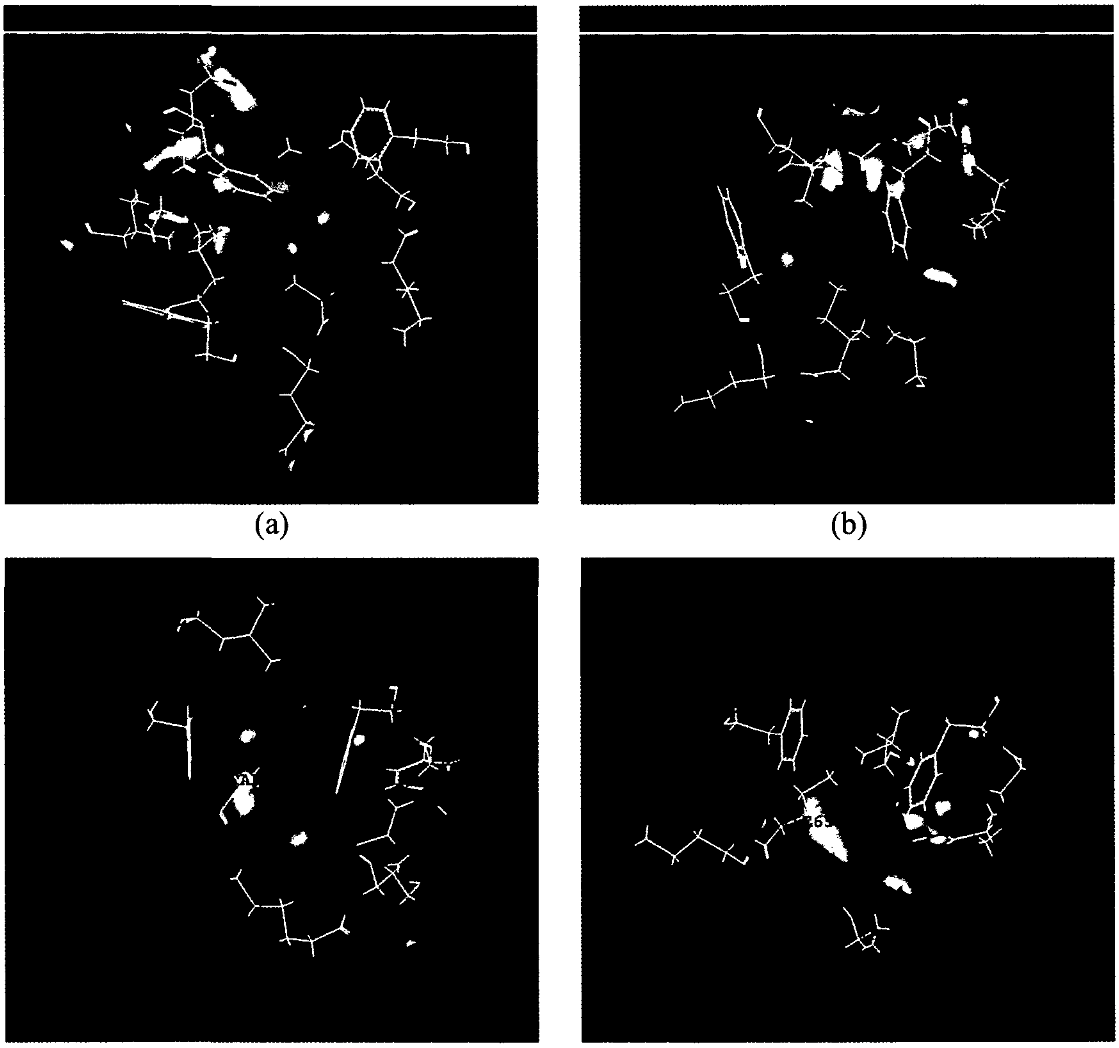


Figure 4-25: Structural conformation of the threonine cavity in SAP and EAT-2.
All four images are presenting the inside of the cavities' surfaces of SAP (left) and EAT-2 (right). In the top pictures (a + b) the different depth of the pockets is illustrated and in the bottom row (c + d) the effect of the cysteine sidechain is shown. From these pictures one can figure out easily how the mercapto-methyl group is limiting the extension of the cavity.

interesting part is again the central pocket. In the middle of the valine cave EAT-2 has only a single shallow hole that is enclosed by a leucine (Leu-93) and isoleucine (Ile-65) residue. SAP has two deeper but smaller cavities at the same position which share a common entrance similar to the entrance of the single EAT-2 hole. These two cavities are encircled by two phenylalanine residues (Phe-77 and Phe-87), one alanine (Ala-66) and one leucine (Leu-43). In contrast to the threonine binding site the valine pockets in SAP and EAT-2 do not contain bound water molecules in the crystal which is due to the hydrophobic character of the residues involved. To illustrate how different the depth of the two pockets actually is, consider that the bottom of the cavity in SAP is formed by Leu-43. This residue corresponds to Leu-42 in EAT-2 which is buried deep inside the protein and does not have any contact to solvent molecules.



(a) (b)

(c) (d)

Figure 4-26: Structural conformation of the valine cavity in SAP and EAT-2.
All four images are presenting the inside of the cavities' surfaces of SAP (left) and EAT-2 (right). The top row shows which residues in both molecules are defining the borders of the cavities. The bottom row shows how Ile-65 and Leu-93 prevent the further extension of the pocket into the inner parts of EAT-2 (d) while the same hole reaches to Leu-43 in SAP (c).

4.3. Elucidating the Phosphatase Activity of SAP

As mentioned in the preceding section the SH2 domain SAP has some unique properties compared to other members of that family. Its binding affinities to signaling peptides of the phosphorylated/dephosphorylated SLAM type are more dependent on the residues upstream of the pTyr than on the C-terminal amino acids [85]. Furthermore SAP is known to block the activity of the SHP-2 phosphatase. Recently, a series of biological experiments by Schweighoffer et. al. [120] discovered that SAP shows a phosphatase activity which cannot be found for other representatives of the SH2 family (such as EAT-2, SHIP, SRC or FYN). This functional similarity to protein-tyrosine phosphatases cannot be verified by a corresponding match of the protein sequences (see Figure 4-27). The problem is thus an interesting test case for the theory that similar protein functions are reflected by similar molecular surfaces [131].

As a reference protein for the similarity searches PTP1, a member of the protein-tyrosine phosphatases (PTPases) family, was selected. These enzymes, in concert with protein-tyrosine kinases, regulate a large number of cellular events, including proliferation and differentiation, metabolism, cytoskeletal organization, neuronal development, and the immune response [67]. PTP1B consists of a single domain which has its active site located at the bottom of a shallow cleft. This site is formed by a sequence of eleven amino acids that represents a common motif of the PTP family and includes the catalytic cysteine and arginine residues. This cysteine residue acts as a nucleophilic agent in the catalytic dephosphorylation reaction.

To elucidate the molecular features that cause the phosphatase activity of SAP the 3D structures of SAP in contact with the peptide fragment SLAM (PDB identifier 1D4W) was compared to an inactive mutant of tyrosine phosphatase PTP1B complexed with bis(para-phosphophenyl)methane, Bppm (PDB identifier 1AAX, see also Figure 4-28) [109]. As can be expected from the low sequence similarity of the two proteins, a direct match between the two structures could not be established by means of the alpha carbon atoms or the protein backbone. However, although the structural features of the two proteins are rather different, the corresponding molecular surfaces around the active sites seem to have similar motifs. Hence a series of surface comparisons between the crystal structures of PTP1B, SAP and EAT-2 (1AAX, 1D4W and 1I3Z respectively) were performed. In these experiments the protein surfaces were restricted to the active sites by selecting only those surface points that were located within 8 Å of the ligands' phosphate groups. For the similarity search these points were then augmented by the ESP of the proteins and the experimental details of that setup are given in Table 4-14. For the sake of simplicity let us assume that the residue 215 in the crystal structure 1AAX is still the

```
SAP     1 -MDAVAVYHGKISRETGEK------LLLATGLDGSYLLRD---------------SESV
            |· ·· |·:|· ·  ::      |:     : ||:| :               ·:·
PTP1B  46 YRDVSPFDHSRIKLHQEDNDYINASLIKMEEAQRSYILTQGPLPNTCGHFWEMVWEQKS

SAP    38 PGVYCL----------CVLYHG------YIYT----------------YRVSQTETGSW
           ||  |          |· |         |:                 | | | |  ·
PTP1B 105 RGVVMLNRVMEKGSLKCAQYWPQKEEKEMIFEDTNLKLTLISEDIKSYYTVRQLELENL

SAP    65 SAETAPGVHKRYFRKIKNLIS----------AFQKPDQGIVIPLQYPVEK---------
          ::: :  : : :: ·  :              |:  :·| : | : ||
PTP1B 164 TTQETREILHFHYTTWPDFGVPESPASFLNFLFKVRESGSLSPEHGPVVVHCSAGIGRS
```

Figure 4-27: Sequence alignment between SAP and PTP1B.
The residues that are in close contact (6.0 Å) to the ligand peptide are highlighted in blue (SAP) and red (PTP1B). A **|** means residue identity and **:**, **·** strong and weak chemical similarity.

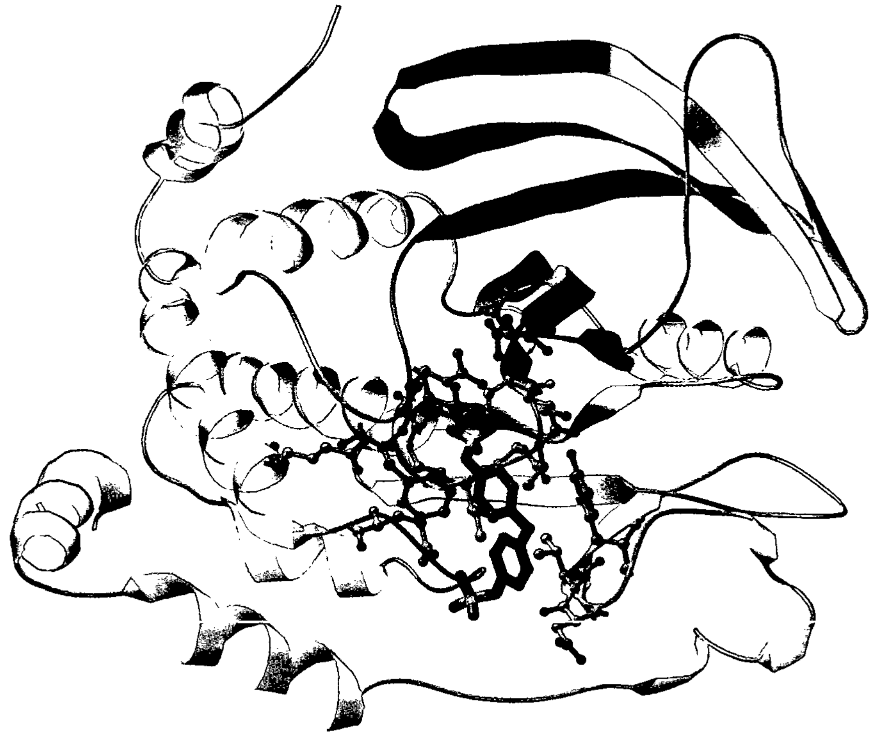


Figure 4-28: Crystal structure of the C215S mutant of PTP1B (1AAX).
The residues of the protein that are within 6.0 Å from the ligand are shown as small ball & sticks. The ligand that is reaching into the active site is rendered with bold ball & sticks. The mutated serine residue is highlighted in red.

natural cysteine.

The investigations discovered a significant surface similarity between the active sites of SAP and PTP1B (see Figure 4-29); it consists of one ridge on one side of the ligands' phenyl rings and two concave patches in the cavity around the phosphate groups of the ligands. In both molecules the phenyl ring of the ligands are surrounded by these similar features and the rest of the cleft that holds them is very well aligned in the superposition of the similar patches. It is interesting to note, that the result of the surface similarity search comes close to an alignment obtained by the plain superposition of the ligands' phenyl rings. No corresponding surface similarity could be detected between the active sites of PTP1B and EAT-2 which correlates well with the biological data.

With the established alignment, one can now look for similar constellations of amino

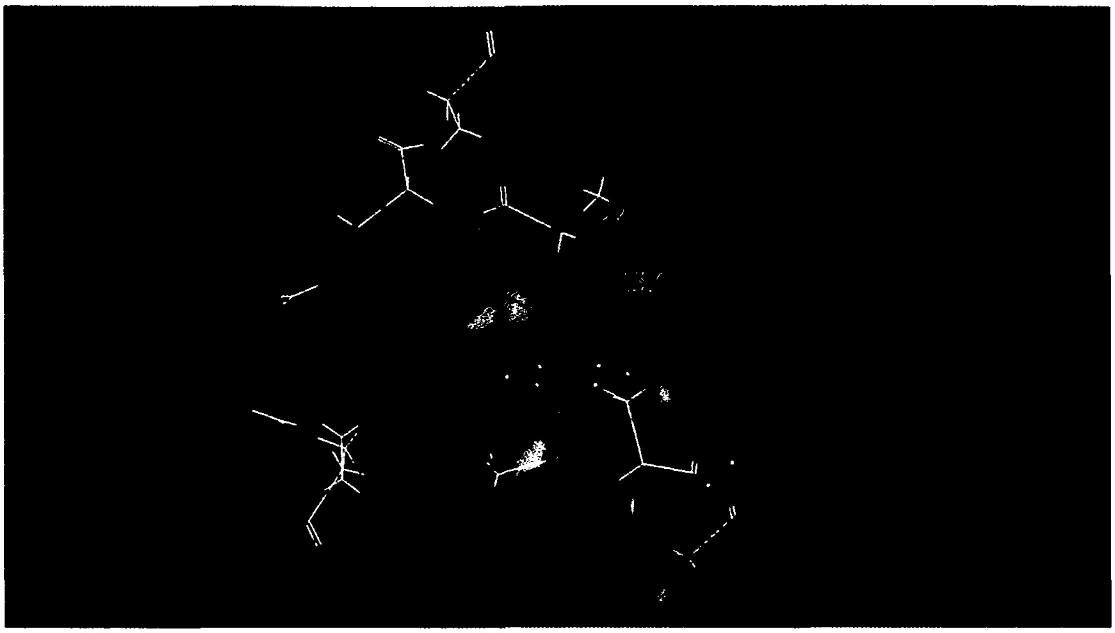
filter parameter	symbol	section[a]	value	property[b]
curvature cut-off range	c_{CR}	2.2.3	2.0 Å	
neighbourhood radius	r_{CP}	3.2	2.0 Å	
fuzzy threshold	F	3.5	0.6	ESP
shape threshold	R	3.6	0.5	STI
distance tolerance	T	3.7	2.0 Å	
minimum distance	δ_{min}	3.7	0.5 Å	
angular tolerance	ϕ_{tol}	3.8	15.0 °	

Table 4-14: Experimental conditions used in the SAP/PTP1B comparisons.
a)the section in the text where the filter is described
b)the molecular surface property applied to the specific filter (ESP, electrostatic potential).

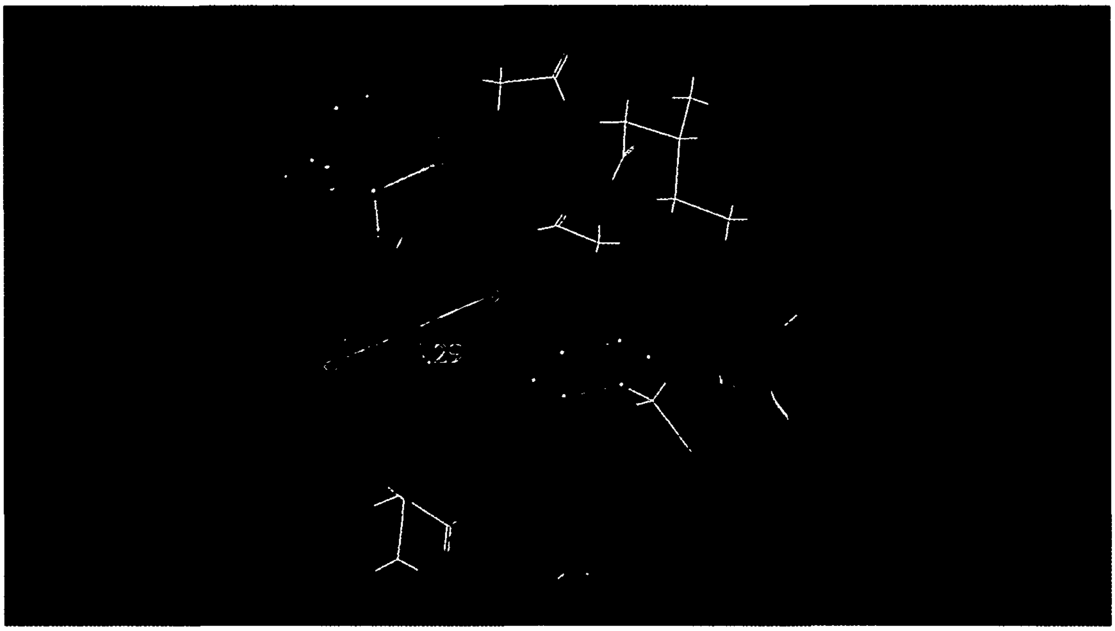
Figure 4-29: Similar surface areas in the active site of PTP1B and SAP.
The similar surfaces in PTP1B (left) and SAP (right) are highlighted in strong colors while the different parts are indicated by less intensive colors. The gray areas are not considered in the surface comparison. The colors are coding the electrostatic potential on the surface; blue represents negative and red positive regions. In both pictures the ligands (Bppm left and pSLAM right) are displayed in balls and sticks with CPK color codes for the elements.

acid residues within the active sites. In both cavities a cysteine and at least one arginine residue are present. These two side chains are involved in the catalytic cleavage of the phosphate group in PTP1B and it is suggested that they are also responsible for the phosphatase activity of SAP. The triangles formed between the cysteine sulfur atom, the central carbon atom of the arginine's guanidine group and the phosphor atom of the ligand are very similar (see Figure 4-30 and Figure 4-31 on page 80). Distances between two atoms in these triangles do not differ by more than 0.5 Å, but the triangles do not coincide in the alignment. However, aligning the triangles would bring the ligands out of a position so that they would not fit into the other active site.

It is suggested that the similar surface regions in both active sites are necessary for the molecular recognition of the ligand structures. In both cases the phenyl ring fits well into the shape of the similar ridge and cleft motif. These structural features may be necessary to bring the substrate in close contact with the catalytic residues. Surface comparison alone cannot answer the question whether the reaction is indeed controlled by the cysteine/arginine residue pairs that are located at different parts of the cleft, because it is a static method that does not consider any dynamic processes. Further structural studies are needed to elucidate the mechanism of the catalytic reaction.



SAP



PTP1B

Figure 4-30: Orientation of the catalytic residues in the active sites.

The residues are displayed to show their orientation with respect to the phosphate groups of the ligands (CPK ball and sticks). The yellow triangle indicates the distances between the ligands' phosphor and the cysteine sulfur and the central carbon of the arginine's guanidine group. The picture clearly shows that the residues in PTP1B (below) are rotated by approximately 180° compared to their counterparts in SAP (above).

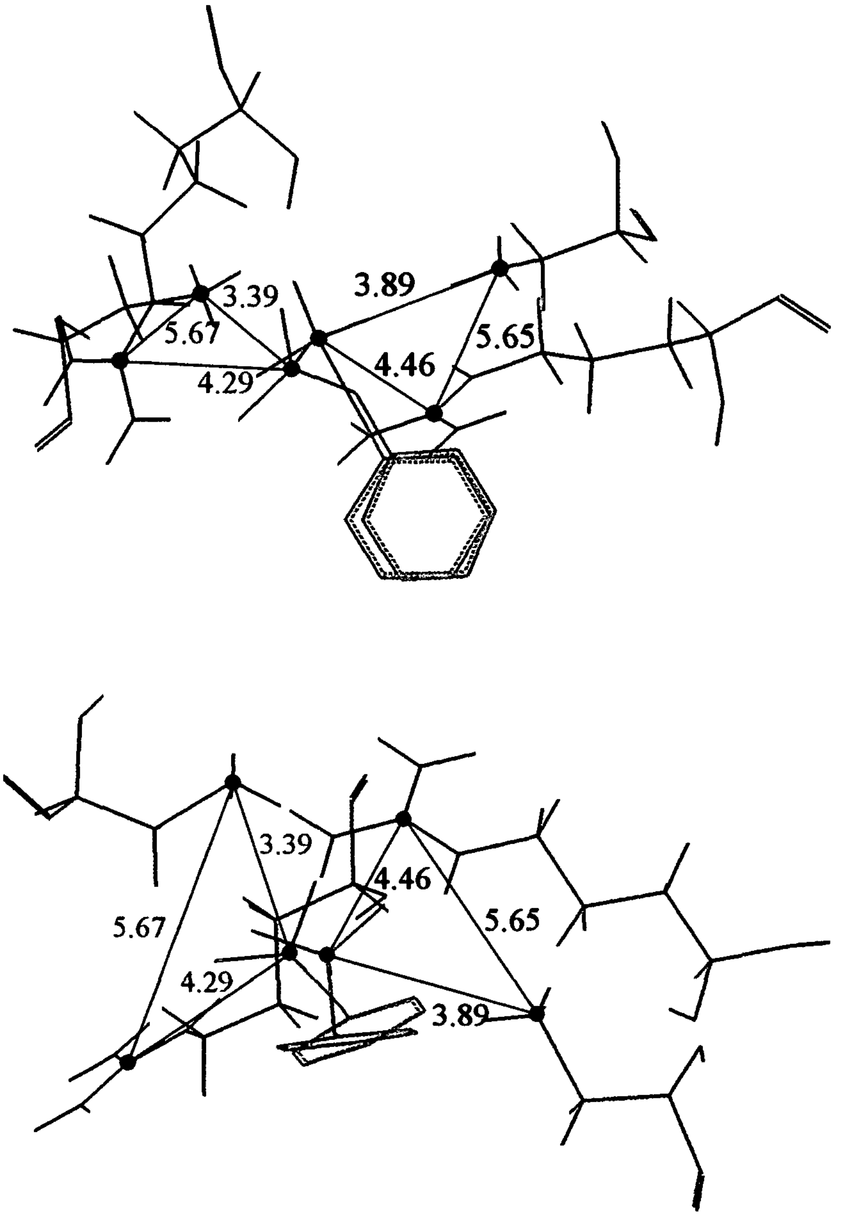


Figure 4-31: Orientation of the catalytic residues and the ligand's phosphate groups
The alignment is based on the surface similarity found between the active sites of SAP and PTP1B. The triangles describe the distances between the important residues and the phosphor atom of the ligands: The distances between the phosphor atoms and the cysteine sulfurs are 3.39 Å (SAP) and 3.89 Å (PTP1B). The carbon atoms of the guanidine group in the arginine residues are placed at distances of 4.46 Å and 4.29 Å and the distance between the two residues is 5.67 Å and 5.65 Å, respectively. The residues of SAP are represented by red and that of PTP1B by blue lines, the phosphate groups and phenyl rings of the ligands are displayed in CPK colors. Top view (above) and side view (below).

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 18th 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.

A The SURFCOMP Program Suite

The SURFCOMP program suite consists of about 30.000 lines of code and performs the heuristic filtering process described in Figure 3-1 together with the preparations of the surfaces and the analysis of the search results. The main program is `surfcomp` which calculates the surface similarities and the functions for the consensus scoring (3.10). The generation and preparation of the molecular surfaces can be performed either via Sybyl 6.91 [2] and MOLCAD [24] or via the MSMS program [114] and a property calculator written in C++. All binary programs developed in this project are available for Linux via source archives or package files in the RPM package manager file format [126]. The evaluation of the results, the ranking, visualization and generation of alignment data, is done by the graphical user interface `surfcomp-monitor` which acts as a plug-in for the geometry viewer Geomview [1].

The input data (the molecular surface objects and data files) and the fundamental comparison parameters are stored and managed by a local or remote MySQL database and scripts are provided that automatically setup and fill the experiment-databases. For the preparation of the surfaces by Sybyl and MOLCAD a suite of SPL scripts is available that provides convenient tools especially for the setup of protein active site comparisons. The suite can handle a series of other surface file formats and is able to calculate basic surface properties like canonical curvatures [141] and the electrostatic potential (section 2.2.1) via the auxiliary program `propgenerator`.

The calls to `surfcomp` are usually invoked via a shell script which is generated from a template by the script `preparesurfcomp`. One can take any user-defined template for that script and it is thereby possible to distribute the single jobs by a scheduler to a Linux cluster or to other high performance computer systems.

A.1 Requirements

The following additional libraries are required

- the Linux operating system
- the Xerces-c XML library (version 2.1 or higher) [124]
- the MySQL client libraries (versions 3.23 or higher)
- Geomview (version 1.8 or higher)
- Perl (version 5.6 or higher) [125]
- A MySQL database server (versions 3.23 or higher) with read/write access for the user.

A.2 Availability

The source code is published under the Novartis open-source license and is available from the web [64] or the attached CD-ROM together with the binary packages for various Linux distributions. Binary executables for other platforms must be created from the source code. Especially compilation for other UNIX compliant systems should be possible with the provided installation tools.

B Publications

J. Chem. Inf. Comput. Sci., **44** (3), 837 -847, 2004.

SURFCOMP: A Novel Graph-Based Approach to Molecular Surface Comparison

Christian Hofbauer, Hans Lohninger, and András Aszódi

Novartis Institutes for BioMedical Research, Brunnerstrasse 59, A-1235 Vienna, Austria, and Institut für Chemische Technologien und Analytik, Technische Universität Wien, Getreidemarkt 9/151, A-1060 Vienna, Austria

Received October 28, 2003

Abstract. Analysis of the distributions of physicochemical properties mapped onto molecular surfaces can highlight important similarities or differences between compound classes, contributing to rational drug design efforts. Here we present an approach that uses maximal common subgraph comparison and harmonic shape image matching to detect locally similar regions between two molecular surfaces augmented with properties such as the electrostatic potential or lipophilicity. The complexity of the problem is reduced by a set of filters that implement various geometric and physicochemical heuristics. The approach was tested on dihydrofolate reductase and thermolysin inhibitors and was shown to recover the correct alignments of the compounds bound in the active sites.

Theory and Applications of Computational Chemistry, Gyeongju, Korea,
14. - 20. Feb. 2004

Molecular surface comparison with SURFCOMP: A novel graph-based approach

Christian Hofbauer, **Hans Lohninger, and András Aszódi**

Institut für Chemische Technologien und Analytik, Technische Universität Wien, Getreidemarkt 9/151, A-1060 Vienna, AUSTRIA
Novartis Institutes for BioMedical Research, Brunnerstrasse 59, A-1235 Vienna, AUSTRIA

Abstract. Analysis of the distributions of physicochemical properties mapped onto molecular surfaces can highlight important similarities or differences between compound classes, contributing to rational drug design efforts [131]. We have developed a method that uses a combination of graph theory, computer vision and computational chemistry to detect local surface similarities between small and medium sized molecules. Our approach is based on 3D structure search where maximal common subgraph isomorphism is used to detect local similarities between the pharmacophoric feature points of different molecules [91]. The extension of this principle to molecular surfaces is cumbersome, because treatment of the complete set of surface points instead of just a few feature points with NP-hard graph algorithms is not feasible. In order to

perform a reliable and fast detection of local surface similarities it is necessary to reduce the complexity of the problem by a set of filters that implement various geometric and physicochemical heuristics.

To achieve this we first generate a simplified representation of the surfaces consisting only of a set of critical points (corresponding to "hills" and "valleys" on the surface), augmented by their surrounding surface patches. Among all possible point pairs we first select those that show sufficient chemical similarity, judged by means of a fuzzy dissimilarity index [48] between physicochemical properties mapped onto the surface points. Then the curvature patterns around all remaining point pairs are compared by harmonic shape image matching [145] to discard points that are not embedded in a similar shape. Finally the distances and angles between combinations of similar pairs are checked to be within certain bounds to form an association graph that is simple enough for the clique detection. The cliques represent the local surface similarities and an alignment between the two molecular surfaces can be calculated based on the corresponding points. Finally the alignments can be clustered to reveal a picture of the total surface similarity between the two molecules.

We tested our method with a dataset of eight thermolysin inhibitors and recovered the correct alignments of the compounds bound in the active sites. The results were in good agreement with another surface-based comparison carried out on the same dataset [37]. We are now directing our efforts to the comparison of protein/protein surfaces and the incorporation of conformational flexibility.

C Abbreviations

ALA	alanine
ARG	arginine
ASP	aspartic acid
ATP	adenosine triphosphate
Bppm	bis(para-phosphophenyl)methane
CP	critical point
CPU	central processing unit
CRK	proto-oncogene C
CYS	cysteine
DHFR	dihydrofolate reductase
DNA	deoxyribonucleic acid
EAT-2	ews/fli1 activated transcript 2
EC	Enzyme Commission
ESP	electrostatic potential
FOL	folic acid
GLU	glutamic acid
GLY	glycine
GRB2	growth factor receptor-bound protein 2
HF	Hartree Fock
HOMO	highest occupied molecular orbital
HSI	harmonic shape image
ILE	isoleucine
LP	lipophilic potential
LUMO	lowest unoccupied molecular orbital
LYS	lysine
MTX	methotrexate
NADP(H)	nicotinamide adenine dinucleotide phosphate
NMR	nuclear magnetic resonance
NOE	nuclear Overhauser effect
NOESY	nuclear Overhauser enhancement Spectroscopy
PDB	Protein Data Bank
PHE	phenylalanine
PTP1B	protein tyrosine phosphatase 1B
PTPases	protein tyrosine phosphatases
pTyr	phosphorylated tyrosine
QSAR	quantitative structure activity relationship
QSD	quadratic shape descriptors
RAM	random access memory
RMS	root mean square

RMSD	root mean square deviation
SAP	SLAM← associated protein
SH2	SRC homology 2
SLAM	signaling lymphocyte activation molecule
SPL	Sybyl programming language
STI	surface topology index
TLN	Thermolysin
TMP	trimethoprim
WRB	Br-WR99210 *(compound name)*

D Indices

D.1 Figure Index

Figure 1-1: ATP binding site of the PDB entry 1B0U (ABC transporter protein). 2
Figure 2-1: Electron density plots parallel to the adenosine ring plane in ATP. 7
Figure 2-2: Creation of solvent accessible and solvent excluded surfaces. 9
Figure 2-3: Different molecular surfaces of ATP. 10
Figure 2-4: Surface autocorrelation function. 15
Figure 2-5: Shape of the membership functions in a linguistic variable. 16
Figure 2-6: A surface patch, its harmonic map and harmonic shape image. 18
Figure 2-7: Border mapping. 19
Figure 2-8: Interpolation scheme for the generation of harmonic shape images 20
Figure 2-9: Grid transformation. 22
Figure 2-10: Pharmacophore match by maximum common subgraph isomorphism. ...23
Figure 3-1: Overview of the surface similarity detection algorithm, developed in this thesis. 26
Figure 3-2: Surface of the thermolysin inhibitor L-valyl-L-tryptophan 28
Figure 3-3: Illustration of the patch generation process. 30
Figure 3-4: Distance filter 33
Figure 3-5: Illustration of the overlap filter. 34
Figure 3-6: Surface of a bacterial ABC transporter protein (PDB: 1L2T). 37
Figure 4-1: The 3D structure of thermolysin (TLN). 42
Figure 4-2: Surface alignment of 1TLP (blue) and 6TMN (red). 44
Figure 4-3: Surface alignment of 1THL (blue) and 1TMN (red). 44
Figure 4-4: Similar surface regions between 1THL (left) and other molecules (right). 46
Figure 4-5: Comparison between the top and closest clusters for 1THL and 1TLP......48
Figure 4-6: Comparison between the top and closest clusters of 1THL and 4TMN. 50
Figure 4-7: Comparison between the top and closest clusters of 1TMN and 4TMN. 50
Figure 4-8: Results of the comparative ranking 52
Figure 4-9: The superposition of 1TLP (blue) and 5TMN (red) 53
Figure 4-10: Alignment of methotrexate and dihydrofolate in the pocket of DHFR. ...56
Figure 4-11: Four different molecular surfaces of the folic acid. 57
Figure 4-12: Line ups between comparisons performed by different surface types......60
Figure 4-13: Alignment of the generated conformations for ATP^{4-}. 62
Figure 4-14: Surface similarity vs. conformational difference. 63
Figure 4-15: Surface alignments of the template and two calc. conformers of ATP^{4-}. .64
Figure 4-16: Alignment of the generated conformations for WRB. 65

Figure 4-17: Separated surface similarities between WRB conformers. 67
Figure 4-18: Structure of the SAP-pSLAM and EAT-2 pSLAM complexes. 68
Figure 4-19: A mechanism for SLAM-induced recruitment and activation of Fyn. 69
Figure 4-20: Relative binding affinities between SAP and different SLAM peptides. 70
Figure 4-21: Sequence alignment between SAP and EAT-2. 70
Figure 4-22: Surface regions considered in the comparison of the SAP and EAT-2. 71
Figure 4-23: Surface differences between SAP (above) and EAT-2 (below). 73
Figure 4-24: Surface differences between SAP and EAT-2 with structures. 74
Figure 4-25: Structural conformation of the threonine cavity in SAP and EAT-2. 75
Figure 4-26: Structural conformation of the valine cavity in SAP and EAT-2. 76
Figure 4-27: Sequence alignment between SAP and PTP1B. 77
Figure 4-28: Crystal structure of the C215S mutant of PTP1B (1AAX). 78
Figure 4-29: Similar surface areas in the active site of PTP1B and SAP. 79
Figure 4-30: Orientation of the catalytic residues in the active sites. 80
Figure 4-31: Orientation of the catalytic residues and the ligand's phosphate groups 81

D.2 Table Index

Table 3-1: Complexity of the association graph 27
Table 3-2: Definition of fuzzy qualitative classes 31
Table 4-1: Experimental conditions used in the thermolysin experiments. 42
Table 4-2: Surface overlays of different thermolysin inhibitors 43
Table 4-3: The number of top level and total alignments 49
Table 4-4: Comparison of closest and top clusters. 49
Table 4-5: Experimental conditions used in the comparative ranking experiments. 51
Table 4-6: Comparative rankings of all molecules of the thermolysin dataset. 55
Table 4-7: Experimental conditions used in the DHFR ligand dataset experiments. 58
Table 4-8: Results obtained for the surface comparison with different surface types. 59
Table 4-9: Results of the surface comparison of ATP^{4-} conformations 63
Table 4-10: Results of the surface comparison between the conformations of WRB. 66
Table 4-11: Experimental conditions used in the conformation tests. 66
Table 4-12: Available protein structures for SAP and EAT-2. 69
Table 4-13: Experimental conditions used in the SAP/EAT-2 comparisons. 72
Table 4-14: Experimental conditions used in the SAP/PTP1B comparisons 78

D.3 Chart Index

Chart 4-1: 2D structures of eight thermolysin inhibitors: 41
Chart 4-2: DHFR inhibitors 56
Chart 4-3: 2D structures of adenosine triphosphate and Br-WR99210 (WBR). 61

E References

(1) Geomview, http://www.geomview.org/ , 1.2.2002.

(2) *SYBYL 6.9*; Tripos Inc.: St. Louis, MO, 2003

(3) POV-Ray the "persistence of vision" raytracer, http://www.povray.org/ , 15.3.2004.

(4) United Devices, http://www.ud.com/home.htm , 2004.

(5) Abagyan, R.; Totrov, M. High-throughput docking for lead generation *Current Opinion in Chemical Biology* **5:4**, 375-382, 2001.

(6) Amann, A.; Bangov, I. P.; Brickmann, J.; Dumitrescu, D.-D.; Klir, G. J.; Mezey, P. G.; Mislow, K.; Rouvray, D. H.; Xu, J. *Fuzzy Logic in Chemistry*; Academic Press: San Diego, CA, 1997.

(7) Appleton, T. Combinatorial chemistry and HTS - feeding a voracious process *Drug Discovery Today* **4:9**, 398-400, 1999.

(8) Aszódi, A. RazorBack 2.0 - linear algebra library, *unpublished work.*

(9) Atkins, P. W.; Friedman, R. S.; Editors. *Molecular Quantum Mechanics, 3rd Edition*; 1996.

(10) Ausiello, G.; Cesareni, G.; Helmer-Citterich, M. ESCHER: a new docking procedure applied to the reconstruction of protein tertiary structure *Proteins: Structure, Function, and Genetics* **28:4**, 556-567, 1997.

(11) Bailey, R. R.; Srinath, M. Orthogonal Moment Features for Use With Parametric and Non-Parametric Classifiers *IEEE Trans.Pattern Anal.Mach.Intell.* **18**, 389-399, 1996.

(12) Barrow, H. G.; Burstall, R. M. Subgraph isomorphism, matching relational structures and maximal cliques *Inf.Process.Lett.* **4:4**, 83-84, 1976.

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

(14) Bhatia, A.; E.áWolf On the circle polynomials of Zernike and related orthogonal sets *Proc.Cambridge.Philosophical.Society.* **50**, 40-48, 1954.

(15) Bladon, P. A rapid method for comparing and matching the spherical parameter surfaces of molecules and other irregular objects *J.Mol.Graphics* **7**, 130-137, 1989.

(16) Blakley, R. L.; Cocco, L. Dismutation of dihydrofolate by dihydrofolate reductase *Biochemistry* **23:11**, 2377-2383, 1984.

(17) Blaney, F.; Flinn, P.; Phippen, R.; Wyatt, R. Molecular surface comparison: Application to drug design *J.Mol.Graphics* **11**, 98-105, 1993.

(18) Blaney, J. M.; Dixon, J. S. A good ligand is hard to find: automated docking methods *Perspect.Drug Discovery Des.* **1:2**, 301-319, 1993.

(19) Boland, M. V.; Markey, M. K.; Murphy, R. F. Automated Recognition of Patterns Characteristic of Subcellular Structures in Fluorescence Microscopy Images *Cytometry* **33**, 366-375, 1998.

(20) Bondi, A. van der Waals volumes and radii *J.Phys.Chem.* **68:3**, 441-451, 1964.

(21) Botfield, M. C.; Green, J. SH2 and SH3 domains: choreographers of multiple signaling pathways *Annual Reports in Medicinal Chemistry* **30**, 227-237, 1995.

(22) Bowen, J. P.; Allinger, N. L. Molecular mechanics: the art and science of parameterization *Rev.Comput.Chem.* **2**, 81-97, 1991.

(23) Brickmann, J.; Bertling, H.; Bussian, B. M.; Goetze, T.; Knoblauch, M.; Waldherr-Teschner, M. MOLCAD - interactive molecular computer graphics on high-performance computers *Tagungsber.- Vortragstag., Ges.Dtsch.Chem., Fachgruppe Chem.-Inf.* **3rd**, 93-111, 1987.

(24) Brickmann, Jürgen, Goetze, Thomas, Heiden, Wolfgang, Moeckel, Gerd, Reiling, Stephan, Vollhardt, Horst, and Zachmann, Carl Dieter Interactive visualization of molecular scenarios with MOLCAD/SYBYL. *Data Visualization Mol. Sci* **1995**,

(25) Brint, A. T.; Willett, P. Algorithms for the identification of three-dimensional maximal common substructures *J.Chem.Inf.Comp.Sci* **27:4**, 152-158, 1987.

(26) Bron, C.; Kerbosch, J. Algorithm 457 - Finding all cliques of an undirected graph *Commun.ACM* **16:9**, 575-577, 1973.

(27) Broughton, H. B. A method for including protein flexibility in protein-ligand docking: improving tools for database mining and virtual screening *Journal of Molecular Graphics & Modelling* **18:3**, 247-257, 2000.

(28) Chan, B.; Lanyi, A.; Song, H. K.; Griesbach, J.; Simarro-Grande, M.; Poy, F.; Howie, D.; Sumegi, J.; Terhorst, C.; Eck, M. J. SAP couples Fyn to SLAM immune receptors *Nat.Cell Biol.* **5:2**, 155-160, 2003.

(29) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins *J.Med.Chem.* **42:25**, 5100-5109, 1999.

(30) Chau, P. L.; Dean, P. M. Molecular recoginition: 3D surface structure comparison by gnomic projection *J.Mol.Graphics* **5:2**, 97-100, 1987.

(31) Chothia, C.; Janin, J. Principles of protein-protein recognition *Nature* **256:5520**, 705-708, 1975.

(32) Connolly, M. L. Analytical molecular surface calculation *J.Appl.Crystallogr.* **16**, 548-558, 1983.

(33) Connolly, M. L. Solvent-accessible surfaces of proteins and nucleic acids *Science* **221:4612**, 709-713, 1983.

(34) Connolly, M. L. Shape complementarity at the hemoglobin α_1 β_1 subunit interface *Biopolymers* **25:7**, 1229-1247, 1986.

(35) Connolly, M. L. Shape distributions of protein topography *Biopolymers* **32:9**, 1215-1236, 1992.

(36) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Mertz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J.Am.Chem.Soc.* **117**, 5179-5197, 1995.

(37) Cosgrove, D.; Bayada, D.; Johnson, A. A novel method of aligning molecules by local surface shape similarity *J.Comput.-Aided Mol.Des.* **14:6**, 573-591, 2000.

(38) Davies, J. F.; Delcamp, T. J.; Prendergast, N. J.; Ashford, V. A.; Freisheim, J. H.; Kraut, J. Crystal structures of recombinant human dihydrofolate reductase complexed with folate and 5-deazafolate *Biochemistry* **29:40**, 9467-9479, 1990.

(39) Desbrun, M., Meyer, M., der, P., and Barr, A. *Discrete differential-geometry operators in nD*,**2000.**

(40) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model *J.Am.Chem.Soc.* **107:13**, 3902-3909, 1985.

(41) Dijkstra, E. A note on two problems in connexion with graphs. *Numerische Mathematik* **1**, 269-271, 1959.

(42) Diller, David J. Homology models, high throughput docking, and drug design. *Abstracts of Papers, 222nd ACS National Meeting, Chicago, IL, United States, August 26-30* **2001**,

(43) Diller, D. J.; Li, R. Kinases, Homology Models, and High Throughput Docking *J.Med.Chem.* **46:22**, 4638-4647, 2003.

(44) Eck, M., DeRose, T., Duchamp, T., Hoppe, H., Lounsbery, M., and Stuetzle, W. *Multiresolution analysis of arbitrary meshes*, University of Washington, Seattle, **1995.**

(45) Eck, M., DeRose, T., Duchamp, T., Hoppe, H., Lounsbery, M., and Stuetzle, W. *Multiresolution analysis of arbitrary meshes*, University of Washington, Seattle, **1995.**

(46) Eells, J.; Sampson, L. Harmonic mappings of Riemannian manifolds *Amer.J.Math.* **86**, 109-160, 1964.

(47) Entzeroth, M. Emerging trends in high-throughput screening *Current opinion in pharmacology* **3:5**, 522-529, 2003.

(48) Exner, T. E.; Keil, M.; Brickmann, J. Pattern recognition strategies for molecular surfaces. I. Pattern generation using fuzzy set theory *Journal of Computational Chemistry* **23:12**, 1176-1187, 2002.

(49) Exner, T. E.; Keil, M.; Brickmann, J. Pattern recognition strategies for molecular surfaces. II. Surface complementarity *Journal of Computational Chemistry* **23:12**, 1188-1197, 2002.

(50) Fastovski, O. High throughput docking: past, present, and future *PharmaChem* **1:7/8**, 18-21, 2002.

(51) Fischer, D.; Lin, S. L.; Wolfson, H. L.; Nussinov, R. A geometry-based suite of molecular docking processes *J.Mol.Biol.* **248:2**, 459-477, 1995.

(52) Forman, M. and Schaffer, P. *Amadeus*, The Saul Zaentz Company, 1984.

(53) Free Software Foundation, The GCC-GNU project, http://gcc.gnu.org/, 27.3.2004.

(54) Ghose, A. K.; Crippen, G. M. Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships. I. Partition coefficients as a measure of hydrophobicity *J.Comp.Chem.* **7:4**, 565-577, 1986.

(55) Goldman, B. B.; Wipke, W. T. QSD quadratic shape descriptors. 2. Molecular docking using quadratic shape descriptors (QSDock) *Proteins: Structure, Function, and Genetics* **38:1**, 79-94, 2000.

(56) Goldman, B. B.; Wipke, W. T. Quadratic Shape Descriptors. 1. Rapid Superposition of Dissimilar Molecules Using Geometrically Invariant Surface Descriptors *J.Chem.Inf.Comp.Sci* **40:3**, 644-658, 2000.

(57) SPDB viewer, http://www.expasy.org/spdbv/, accessed on 2004.

(58) Guntert, P. Automated NMR protein structure calculation *Progress in nuclear magnetic resonance spectroscopy* **43:3-4**, 105, 2003.

(59) Heiden, W.; Moeckel, G.; Brickmann, J. A new approach to analysis and display of local lipophilicity/hydrophilicity mapped on molecular surfaces *J.Comput.-Aided Mol.Des.* **7:5**, 503-514, 1993.

(60) Heiden, W.; Brickmann, J. Segmentation of protein surfaces using fuzzy logic *J.Mol.Graphics* **12:2**, 106-115, 1994.

(61) Hekmat-Nejad, M.; Rathod, P. K. Plasmodium falciparum: kinetic interactions of WR99210 with pyrimethamine-sensitive and pyrimethamine-resistant dihydrofolate reductase *Exp.Parasitol.* **87:3**, 222-228, 1997.

(62) Hew, Patrick C. and Alder, Michael D. *Zernike or orthogonal Fourier-Mellon Moments for representing and recognising printed digits*, Department of Mathematics, The University of Western Australia, **1998.**

(63) Hoey, J.; Little, J. J. Representation and recognition of complex human motion; In *Conference on Computer Vision and Pattern Recognition*; 2000; pp 752-759.

(64) Hofbauer, C., The SURFCOMP program suite, http://teachme.tuwien.ac.at/surfcomp , 2004.

(65) Hofbauer, C.; Lohninger, H.; Aszódi, A. SURFCOMP: A novel graph-based approach to molecular surface comparison *J.Chem.Inf.Comp.Sci* **44:3**, 837-847, 2004.

(66) Humphrey, W.; Dalke, A.; Schulten, K. VMD - Visual Molecular Dynamics *J.Mol.Graphics* **14**, 33-38, 1996.

(67) Hunter, T. Protein kinases and phosphatases: the yin and yang of protein phosphorylation and signaling *Cell* **80:2**, 225-236, 1995.

(68) Hwang, P. M.; Li, C.; Morra, M.; Lillywhite, J.; Muhandiram, D. R.; Gertler, F.; Terhorst, C.; Kay, L. E.; Pawson, T.; Forman-Kay, J. D.; Li, S. C. A "three-pronged" binding mechanism for the SAP/SH2D1A SH2 domain: structural basis and relevance to the XLP syndrome *EMBO J.* **21:3**, 314-323, 2002.

(69) Inglese, J. Expanding the HTS paradigm *Drug Discovery Today* **7:18**, S105-S106, 2002.

(70) International Union of Biochemistry *Enzyme Nomenclature 1992*; Academic Press: New York, 1993.

(71) Karp, R. M. Reducibility among combinatorial problems; In *Complexity of Computer Computations*; Miller, R. E., Thatcher, J. W., eds. Plenum Press: New York, 1972; pp 85-103.

(72) Katchalski-Katzir, E.; Shariv, I.; Eisenstein, M.; Friesem, A. A.; Aflalo, C.; Vakser, I. A. Molecular surface recognition: determination of geometric fit

between proteins and their ligands by correlation techniques *Proceedings of the National Academy of Sciences of the United States of America* **89:6**, 2195-2199, 1992.

(73) Keil, M. *Modellierung und Vorhersage von Strukturen biomolekularer Assoziate auf der Basis von statistischen Datenbankanalysen.* PhD Thesis, Technische Universität Darmstadt, Darmstadt, **2002**.

(74) Khotanzad, A.; Hong, Y. H. Invariant image recognition by Zernike moments *IEEE Trans.Pattern Anal.Mach.Intell.* **12:5**, 489-497, 1990.

(75) Koshland, D. E., Jr. Correlation of structure and function in enzyme action *Science* **142:3599**, 1533-1541, 1963.

(76) Koshland, D. E., Jr. Role of flexibility in the specificity, control and evolutiion of enzymes *FEBS letters* **62**, 47-52, 1976.

(77) Lamdan, Y. and Wolfson, H. J. Geometric Hashing: A General and Efficient Model Based Recognition Scheme. *International Conference on Computer Vision* **1988**, 213-229.

(78) Latour, S.; Roncagalli, R.; Chen, R.; Bakinowski, M.; Shi, X.; Schwartzberg, P. L.; Davidson, D.; Veillette, A. Binding of SAP SH2 domain to FynT SH3 domain reveals a novel mechanism of receptor signalling in immune regulation *Nat.Cell Biol.* **5:2**, 149-154, 2003.

(79) Leach, A. R. *Molecular Modelling. Principles and Applications.*; Prentice Hall: Upper Saddle River, NJ, USA, 2001.

(80) Lee, B.; Richards, F. M. The interpretation of protein structures: estimation of static accessibility *J.Mol.Biol.* **55**, 379-400, 1971.

(81) Leicester, S. E.; Finney, J. L.; Bywater, R. P. Description of molecular surface shape using Fourier descriptors *J.Mol.Graphics* **6:2**, 104-8, 100, 1988.

(82) Lemmen, C.; Lengauer, T. Time-efficient flexible superposition of medium-sized molecules *J.Comput.-Aided Mol.Des.* **11:4**, 357-368, 1997.

(83) Li, C.; Iosef, C.; Jia, C. Y.; Han, V. K.; Li, S. S. Dual functional roles for the X-linked lymphoproliferative syndrome gene product SAP/SH2D1A in signaling through the signaling lymphocyte activation molecule (SLAM) family of immune receptors *J.Biol.Chem.* **278:6**, 3852-3859, 2003.

(84) Li, R.; Sirawaraporn, R.; Chitnumsub, P.; Sirawaraporn, W.; Wooden, J.; Athappilly, F.; Turley, S.; Hol, W. G. Three-dimensional structure of m. tuberculosis dihydrofolate reductase reveals opportunities for the design of novel tuberculosis drugs *J.Mol.Biol.* **295:2**, 307-323, 2000.

(85) Li, S. C.; Gish, G.; Yang, D.; Coffey, A. J.; Forman-Kay, J. D.; Ernberg, I.; Kay, L. E.; Pawson, T. Novel mode of ligand binding by the SH2 domain of the human XLP disease gene product SAP/SH2D1A *Curr.Biol.* **9:23**, 1355-1362, 1999.

(86) Lin, S. L.; Nussinov, R.; Fischer, D.; Wolfson, H. J. Molecular surface representations by sparse critical points *Prot.Struct.Func.Gen.* **18:1**, 94-101, 1994.

(87) Little, J. J., Hoey, J., and Boyd, J. Characterizing 2D Flow Fields with Zernike, *unpublished work.*

(88) Maclennan, A. J.; Shaw, G. A yeast SH2 domain *Trends Biochem.Sci.* **18:12**, 464-465, 1993.

(89) Markey, M. K.; Boland, M. V.; Murphy, R. F. Toward Objective Selection of Representative Microscope Images *Biophys.J.* **76**, 2230-2237, 1999.

(90) Marshall, G. R.; Barry, C. D.; Bosshard, H. E.; Dammkoehler, R. A.; Dunn, D. A. The conformational parameter in drug design: the active analog approach *ACS Symposium Series* **112.**, 205-226, 1979.

(91) Martin, Y. C.; Bures, M. G.; Danaher, E. A.; DeLazzer, J.; Lico, I.; Pavlik, P. A. A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists *J.Comput.-Aided Mol.Des.* **7:1**, 83-102, 1993.

(92) Masek, B. B. Molecular surface comparisons. *Molecular Similarity in Drug Design* **1995**,

(93) McLachlan, A. Gene duplications in the structural evolution of chymotrypsin *J.Mol.Biol.* **128:1**, 49-79, 1979.

(94) McPherson, J. D.; Marra, M.; Hillier, L.; Waterston, R. H.; Chinwalla, A.; Wallis, J.; Sekhon, M.; Wylie, K.; Mardis, E. R.; Wilson, R. K.; Fulton, R.; Kucaba, T. A.; Wagner-McPherson, C.; Barbazuk, W. B.; Gregory, S. G.; Humphray, S. J.; French, L.; Evans, R. S.; Bethel, G.; Whittaker, A.; Holden, J. L.; McCann, O. T.; Dunham, A.; Soderlund, C.; Scott, C. E.; Bentley, D. R.; Schuler, G.; Chen, H. C.; Jang, W.; Green, E. D.; Idol, J. R.; Maduro, V. V.; Montgomery, K. T.; Lee, E.; Miller, A.; Emerling, S.; Kucherlapati; Gibbs, R.; Scherer, S.; Gorrell, J. H.; Sodergren, E.; Clerc-Blankenburg, K.; Tabor, P.; Naylor, S.; Garcia, D.; de Jong, P. J.; Catanese, J. J.; Nowak, N.; Osoegawa, K.; Qin, S.; Rowen, L.; Madan, A.; Dors, M.; Hood, L.; Trask, B.; Friedman, C.; Massa, H.; Cheung, V. G.; Kirsch, I. R.; Reid, T.; Yonescu, R.; Weissenbach, J.; Bruls, T.; Heilig, R.; Branscomb, E.; Olsen, A.; Doggett, N.; Cheng, J. F.; Hawkins, T.; Myers, R. M.; Shang, J.; Ramirez, L.; Schmutz, J.; Velasquez, O.; Dixon, K.; Stone, N. E.; Cox, D. R.; Haussler, D.; Kent, W. J.; Furey, T.; Rogic, S.; Kennedy, S.; Jones, S.; Rosenthal, A.; Wen, G.; Schilhabel, M.; Gloeckner, G.; Nyakatura, G.; Siebert, R.; Schlegelberger, B.; Korenberg, J.; Chen, X. N.; Fujiyama, A.; Hattori, M.; Toyoda, A.; Yada, T.; Park, H. S.; Sakaki, Y.; Shimizu, N.; Asakawa, S.; Kawasaki, K.; Sasaki, T.; Shintani, A.; Shimizu, A.; Shibuya, K.; Kudoh, J.; Minoshima, S.; Ramser, J.; Seranski, P.; Hoff, C.; Poustka, A.; Reinhardt, R.; Lehrach, H. A physical map of the human genome *Nature* **409:6822**, 934-941, 2001.

(95) Meza, M. B. Bead-based HTS applications in drug discovery *Drug Discovery Today* **5:Supplement 1**, 38-41, 2000.

(96) Mezey, P. G. The degree of similarity of three-dimensional bodies: application to molecular shape analysis *J.Math.Chem.* **7:1**, 39-49, 1991.

(97) MichaelLounsbery, J. W. Multiresolution analysis for surface of arbitrary topological type *ACM.Transactions.on.Graphics.* **16**, 34-73, 1997.

(98) Mirau, P. A.; Heffner, S. A.; Bovey, F. A. Three-dimensional nuclear Overhauser effect/J-resolved spectroscopy *Journal of Magnetic Resonance (1969-1992)* **89:3**, 572-577, 1990.

(99) Morra, M. Structural basis for the interaction of the free SH2 domain EAT-2 with SLAM receptors in hematopoietic cells *The EMBO journal* **20:21**, 5840, 2001.

(100) Morra, M.; Howie, D.; Grande, M. S.; Sayos, J.; Wang, N.; Wu, C.; Engel, P.; Terhorst, C. X-linked lymphoproliferative disease: a progressive immunodeficiency *Annu.Rev.Immunol.* **19**, 657-682, 2001.

(101) Mukundan, R.; Ramakrishnan, K. R. *Moment functions in image analysis*; World Scientific: Singapore, 1998.

(102) Norel, R.; Lin, S. L.; Wolfson, H. J.; Nussinov, R. Shape complementarity at protein-protein interfaces *Biopolymers* **34:7**, 933-940, 1994.

(103) Norel, R.; Lin, S. L.; Wolfson, H. J.; Nussinov, R. Molecular surface complementarity at protein-protein interfaces: the critical role played by surface normals at well placed, sparse, points in docking *J.Mol.Biol.* **252:2**, 263-273, 1995.

(104) Perutz, M. F. Structure of hemoglobin *Brookhaven.Symp.Biol.* **13**, 165-183, 1960.

(105) Polanski, J.; Gasteiger, J.; Wagener, M.; Sadowski, J. The comparison of molecular surfaces by neural networks and its applications to quantitative structure activity studies *Quant.Struct.-Act.Relat.* **17:1**, 27-36, 1998.

(106) Polshakov, V. I.; Morgan, W. D.; Birdsall, B.; Feeney, J. Validation of a new restraint docking method for solution structure determinations of protein-ligand complexes *Journal of Biomolecular NMR* **14:2**, 115-122, 1999.

(107) Poy, F.; Yaffe, M. B.; Sayos, J.; Saxena, K.; Morra, M.; Sumegi, J.; Cantley, L. C.; Terhorst, C.; Eck, M. J. Crystal structures of the XLP protein SAP reveal a class of SH2 domains with extended, phosphotyrosine-independent sequence recognition *Mol.Cell* **4:4**, 555-561, 1999.

(108) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. LU Decomposition and Its Applications.; In *Numerical Recipes in FORTRAN: The Art of Scientific Computing*; Cambridge University Press: Cambridge, England, 1992; pp 34-42.

(109) Puius, Y. A.; Zhao, Y.; Sullivan, M.; Lawrence, D. S.; Almo, S. C.; Zhang, Z. Y. Identification of a second aryl phosphate-binding site in protein-tyrosine phosphatase 1B: a paradigm for inhibitor design *Proc.Natl.Acad.Sci.U.S.A* **94:25**, 13420-13425, 1997.

(110) Raevsky, O. A.; Grigor'ev, V. Yu.; Kireev, D. B.; Zefirov, N. S. Complete Thermodynamic Description of H-Bonding in the Framework of Multiplicative Approach *Quant.Struct.-Act.Relat.* **11**, 49-63, 1992.

(111) Raevsky, O. A.; Skvortsov, V. S. 3D hydrogen bond thermodynamics (HYBOT) potentials in molecular modelling *J.Comput.-Aided Mol.Des.* **16:1**, 1-10, 2002.

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

(113) Ritchie, D. W. A. Protein docking using spherical polar Fourier correlations *Prot.Struct.Func.Gen.* **39**, 178-194, 2000.

(114) Sanner, M. F., Python based software developments, *http://www.scripps.edu/~sanner/python/index.html*, 14.1.2004.

(115) Sanner, M. F., Olson, A. J., and Spehner, J. Fast and robust computation of molecular surfaces. *Proc. 11th ACM Symp. Comp. Geom* **1995**, C6-C7.

(116) Saunders, M. Stochastic exploration of molecular mechanics energy surfaces. Hunting for the global minimum *J.Am.Chem.Soc.* **109:10**, 3150-3152, 1987.

(117) Sawyer, T. K. Src homology-2 domains: structure, mechanisms, and drug discovery *Biopolymers* **47:3**, 243-261, 1998.

(118) Schaffhausen, B. SH2 domain structure and function *Biochim.Biophys.Acta* **1242:1**, 61-75, 1995.

(119) Schrödinger Inc., Schrödinger: Jaguar Program, *http://www.schrodinger.com/Products/jaguar.html*, 29.3.2004.

(120) Schweighoffer, T. Investigating the Phosphatase Activities of SH2 Domains, *unpublished work.*

(121) Schweitzer, B. I.; Dicker, A. P.; Bertino, J. R. Dihydrofolate reductase as a therapeutic target *FASEB J.* **4:8**, 2441-2452, 1990.

(122) Shoichet, B. K.; Kuntz, I. D. Matching chemistry and shape in molecular docking *Protein Eng.* **6:7**, 723-732, 1993.

(123) Siek, J., Lee, L.-Q., and Lumsdaine, A., The Boost graph library, *http://www.boost.org/libs/graph/doc/index.html*, 2004.

(124) The Apache Software Foundation, Xerces C++ Parser, *http://xml.apache.org/xerces-c/index.html*, 2004.

(125) The Perl Foundation, The Perl Directory, *http://www.perl.org/*, 2004.

(126) the RPM community, RPM Package Manager, *http://www.rpm.org/*, 2004.

(127) van der Waals, J. D. *Die Zustandsgleichung*;

(128) van der Waals, J. D. The volume of the molecule and the volume of the constituent atoms *Verslag van de Gewone Vergadering van de Afdeling Natuurkunde, Koninklijke Nederlandse Akademie van Wetenschappen* **22**, 782-792, 1914.

(129) Vangrevelinghe, E.; Zimmermann, K.; Schoepfer, J.; Portmann, R.; Fabbro, D.; Furet, P. Discovery of a Potent and Selective Protein Kinase CK2 Inhibitor by High-Throughput Docking *J.Med.Chem.* **46:13**, 2656-2662, 2003.

(130) Venter, J. C.; Adams, M. D.; Myers, E. W.; Li, P. W.; Mural, R. J.; Sutton, G. G.; Smith, H. O.; Yandell, M.; Evans, C. A.; Holt, R. A.; Gocayne, J. D.; Amanatides, P.; Ballew, R. M.; Huson, D. H.; Wortman, J. R.; Zhang, Q.; Kodira, C. D.; Zheng, X. H.; Chen, L.; Skupski, M.; Subramanian, G.; Thomas, P. D.; Zhang, J.; Gabor Miklos, G. L.; Nelson, C.; Broder, S.; Clark, A. G.; Nadeau, J.; McKusick, V. A.; Zinder, N.; Levine, A. J.; Roberts, R. J.; Simon, M.; Slayman, C.; Hunkapiller, M.; Bolanos, R.; Delcher, A.; Dew, I.; Fasulo, D.; Flanigan, M.; Florea, L.; Halpern, A.; Hannenhalli, S.; Kravitz, S.; Levy, S.; Mobarry, C.; Reinert, K.; Remington, K.; Abu-Threideh, J.; Beasley, E.; Biddick, K.; Bonazzi, V.; Brandon, R.; Cargill, M.; Chandramouliswaran, I.; Charlab, R.; Chaturvedi, K.; Deng, Z.; Di, F., V; Dunn, P.; Eilbeck, K.; Evangelista, C.; Gabrielian, A. E.; Gan, W.; Ge, W.; Gong, F.; Gu, Z.; Guan, P.; Heiman, T. J.; Higgins, M. E.; Ji, R. R.; Ke, Z.; Ketchum, K. A.; Lai, Z.; Lei, Y.;

Li, Z.; Li, J.; Liang, Y.; Lin, X.; Lu, F.; Merkulov, G. V.; Milshina, N.; Moore, H. M.; Naik, A. K.; Narayan, V. A.; Neelam, B.; Nusskern, D.; Rusch, D. B.; Salzberg, S.; Shao, W.; Shue, B.; Sun, J.; Wang, Z.; Wang, A.; Wang, X.; Wang, J.; Wei, M.; Wides, R.; Xiao, C.; Yan, C.; Yao, A.; Ye, J.; Zhan, M.; Zhang, W.; Zhang, H.; Zhao, Q.; Zheng, L.; Zhong, F.; Zhong, W.; Zhu, S.; Zhao, S.; Gilbert, D.; Baumhueter, S.; Spier, G.; Carter, C.; Cravchik, A.; Woodage, T.; Ali, F.; An, H.; Awe, A.; Baldwin, D.; Baden, H.; Barnstead, M.; Barrow, I.; Beeson, K.; Busam, D.; Carver, A.; Center, A.; Cheng, M. L.; Curry, L.; Danaher, S.; Davenport, L.; Desilets, R.; Dietz, S.; Dodson, K.; Doup, L.; Ferriera, S.; Garg, N.; Gluecksmann, A.; Hart, B.; Haynes, J.; Haynes, C.; Heiner, C.; Hladun, S.; Hostin, D.; Houck, J.; Howland, T.; Ibegwam, C.; Johnson, J.; Kalush, F.; Kline, L.; Koduru, S.; Love, A.; Mann, F.; May, D.; McCawley, S.; McIntosh, T.; McMullen, I.; Moy, M.; Moy, L.; Murphy, B.; Nelson, K.; Pfannkoch, C.; Pratts, E.; Puri, V.; Qureshi, H.; Reardon, M.; Rodriguez, R.; Rogers, Y. H.; Romblad, D.; Ruhfel, B.; Scott, R.; Sitter, C.; Smallwood, M.; Stewart, E.; Strong, R.; Suh, E.; Thomas, R.; Tint, N. N.; Tse, S.; Vech, C.; Wang, G.; Wetter, J.; Williams, S.; Williams, M.; Windsor, S.; Winn-Deen, E.; Wolfe, K.; Zaveri, J.; Zaveri, K.; Abril, J. F.; Guigo, R.; Campbell, M. J.; Sjolander, K. V.; Karlak, B.; Kejariwal, A.; Mi, H.; Lazareva, B.; Hatton, T.; Narechania, A.; Diemer, K.; Muruganujan, A.; Guo, N.; Sato, S.; Bafna, V.; Istrail, S.; Lippert, R.; Schwartz, R.; Walenz, B.; Yooseph, S.; Allen, D.; Basu, A.; Baxendale, J.; Blick, L.; Caminha, M.; Carnes-Stine, J.; Caulk, P.; Chiang, Y. H.; Coyne, M.; Dahlke, C.; Mays, A.; Dombroski, M.; Donnelly, M.; Ely, D.; Esparham, S.; Fosler, C.; Gire, H.; Glanowski, S.; Glasser, K.; Glodek, A.; Gorokhov, M.; Graham, K.; Gropman, B.; Harris, M.; Heil, J.; Henderson, S.; Hoover, J.; Jennings, D.; Jordan, C.; Jordan, J.; Kasha, J.; Kagan, L.; Kraft, C.; Levitsky, A.; Lewis, M.; Liu, X.; Lopez, J.; Ma, D.; Majoros, W.; McDaniel, J.; Murphy, S.; Newman, M.; Nguyen, T.; Nguyen, N.; Nodell, M. The sequence of the human genome *Science* **291:5507**, 1304-1351, 2001.

(131) Via, A.; Ferrè, F.; Brannetti, B.; Helmer-Citterich, M. Protein surface similarities: A survey of methods to describe and compare protein surfaces *Cell.Mol.Life Sci.* **57**, 1970-1977, 2000.

(132) Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of molecular surface properties for modeling corticosteroid binding globulin and cytosolic Ah receptor activity by neural networks *J.Am.Chem.Soc.* **117:29**, 7769-7775, 1995.

(133) Wang, H. Grid-search molecular accessible surface algorithm for solving the protein docking problem *Journal of Computational Chemistry* **12:6**, 746-750, 1991.

(134) Wang, J.; Hou, T.; Chen, L.; Xu, X. Automated docking of peptides and proteins by genetic algorithm *Chemometrics and Intelligent Laboratory Systems* **45:1,2**, 281-286, 1999.

(135) Wang, R.; Wang, S. How Does Consensus Scoring Work for Virtual Library Screening? An Idealized Computer Experiment *J.Chem.Inf.Comp.Sci* **41:5**, 1422, 2001.

(136) Watson, J. D.; Crick, F. H. The structure of DNA *Cold Spring Harb.Symp.Quant.Biol.* **18**, 123-131, 1953.

(137) Welch, W.; Ruppert, J.; Jain, A. N. Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites *Chemistry & Biology* **3:6**, 449-462, 1996.

(138) Whitley, D. C. Van der Waals surface graphs and molecular shape *J.Math.Chem.* **23**, 377-397, 1998.

(139) Wüthrich, K. Three-dimensional structures of noncrystalline proteins observed by nuclear magnetic resonance *Chemica Scripta* **29A**, 23-26, 1989.

(140) Yang, J. M.; Kao, C. Y. Flexible ligand docking using a robust evolutionary algorithm *Journal of Computational Chemistry* **21:11**, 988-998, 2000.

(141) Zachmann, C. D.; Heiden, W.; Schlenkrich, M.; Brickmann, J. Topological analysis of complex molecular surfaces *J.Comp.Chem.* **13:1**, 76-84, 1992.

(142) Zadeh, L. A. Fuzzy Sets *Inform.Control.* **8**, 338-353, 1965.

(143) Zernike, F. Beugungstheorie des Schneidenverfahrens und seiner verbesserten Form, der Phasenkontrastmethode *Physica.* **1**, 689-704, 1934.

(144) Zhang, D. *Harmonic Shape Images: A 3D free-form surface representation and its applications in surface matching.* PhD Thesis, Carnegie Mellon University, Pittsburgh, **1999**.

(145) Zhang, D. and Herbert, M. Harmonic Maps and their applications in surface matching. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR '99)* **1999**,

Lebenslauf

Angaben zur Person:

Name:	Dipl.-Ing. Christian HOFBAUER
Adresse:	Gentzgasse 15/2/28, A-1180 Wien, Österreich
Telefon/Faxnr.:	++43-(1)-968 85 03
e-mail:	chofbaue@qspr03.tuwien.ac.at
Geburtstag und -ort:	13. August 1976, Eisenstadt
Staatsbürgerschaft:	Österreich
Geschlecht:	männlich
Familienstand:	ledig

Ausbildung:

2001-2004:	Dissertation am Novartis Institute of Biomedical Research in Wien (Rigorosum geplant für Okt. 2004)
1995-2001:	Technische Universität, Wien Studienrichtung Technische Chemie, Studienzweig Organische Chemie und Technologie
1994:	Matura mit ausgezeichnetem Erfolg
1986 – 1994:	Albertus Magnus Gymnasium, Wien
1982 – 1986:	Volksschule, Wien

Zusätzliche Ausbildung:

- Projektmanagementkurse an der Universität für Bodenkultur und an der Technischen Universität Wien
- Einführungsvorlesungen zu Software Engineering (Rational Unified Process, UML), und „Space Based Computing“ an der Technischen Universität Wien
- sehr gute Kenntnisse in folgenden Programmiersprachen: C++, Java, Pascal, Perl, PHP, Python, Visual Basic for Applications
- fortgeschrittene Kenntnisse in Administration von Windows und Linux Systemen, HTML, Office Anwendungen und GUI Entwicklung.
- Grundlagenwissen in Elektrotechnik

Sprachkenntnisse:

Muttersprache: Deutsch

Sprache	**Sprechen**	**Verstehen**	**Lesen**	**Schreiben**
Englisch	sehr gut	sehr gut	sehr gut	sehr gut

Berufserfahrung:

02-2004 - 04-2004: Lektor an der Fachhochschule Wiener Neustadt: Veranstaltung eines Borland Delphi Programmierkurses gemeinsam mit Prof. Dr. Hans Lohninger.

11-2001 - 10-2004: Novartis Institute of Biomedical Research, Wien: Wissenschaftlicher Mitarbeiter in der „In Silico Sciences" Unit unter Dr. András Aszódi.

02-2000 - 10-2001: Institut für Analytische Chemie, Arbeitsgruppe von Prof. Dr. Hans Lohninger:
Projektkoordination für das elektronische Lehrbuch "Teach/Me Instrumentelle Analytik".

08-1999: Ferialpraktikum am Novartis Forschungsinstitut in Wien - im Labor von Dr. Jan-Markus Seifert - auf dem Gebiet der kombinatorischen, oganischen Synthese.

08-1998: Ferialpraktikant bei IBM Österreich als Mitarbeiter im Projektbüro eines Y2K Projektes bei einer großen Österreichischen Versicherung.

07-1997: Ferialpraktikum im Betriebslabor der Agrana Zuckerfabrik in Tulln.

08-1996: Ferialpraktikum im Rechenzentrum der Austrian Research Centers in Seibersdorf.

Publikationen:

C. Hofbauer, H. Lohninger, A. Aszódi SURFCOMP: A Novel Graph-Based Approach to Molecular Surface Comparison. *J. Chem. Inf. Comput. Sci.* **44:3**, 837-847, 2004.