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:

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

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:

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

where **r** is the interatomic distance and *e* as well as *s* 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:

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

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, $r(\mathbf{r})$, 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:

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

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

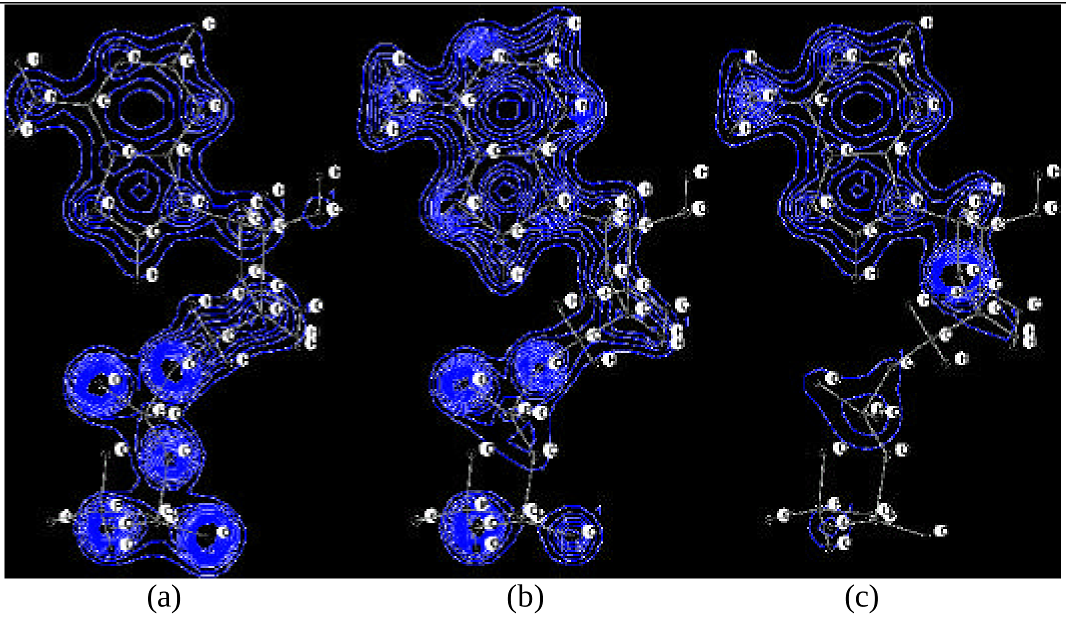


Figure 2-1: Electron density plots parallel to the adenosine ring plane in ATP.
(a) 1.0 Å below, (b) at and (c) 1.0 Å above the ring plane.

of the atoms (see eq. 2-5). The function parameters c and a 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:

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

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

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

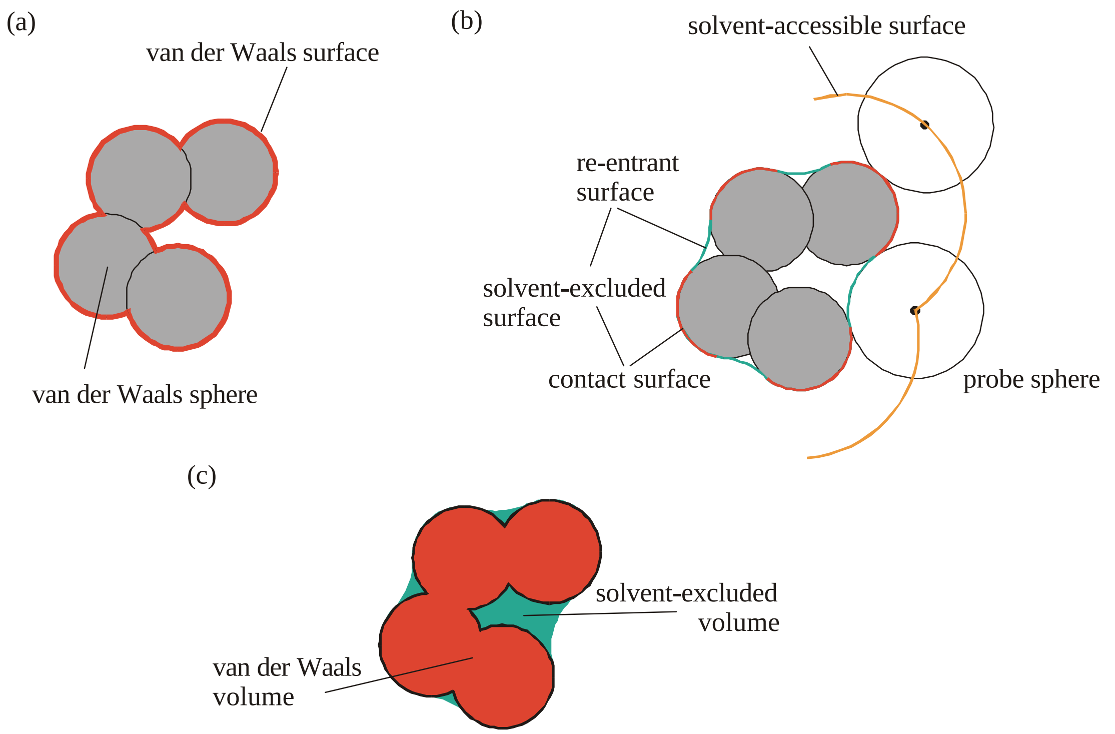


Figure 2-2: Creation of solvent accessible and solvent excluded surfaces.
(a) van der Waals surface (b) creation of the solvent accessible and solvent excluded surface and (c) comparison of the van der Waals and solvent excluded volume. Picture (b) is redrawn from [79].

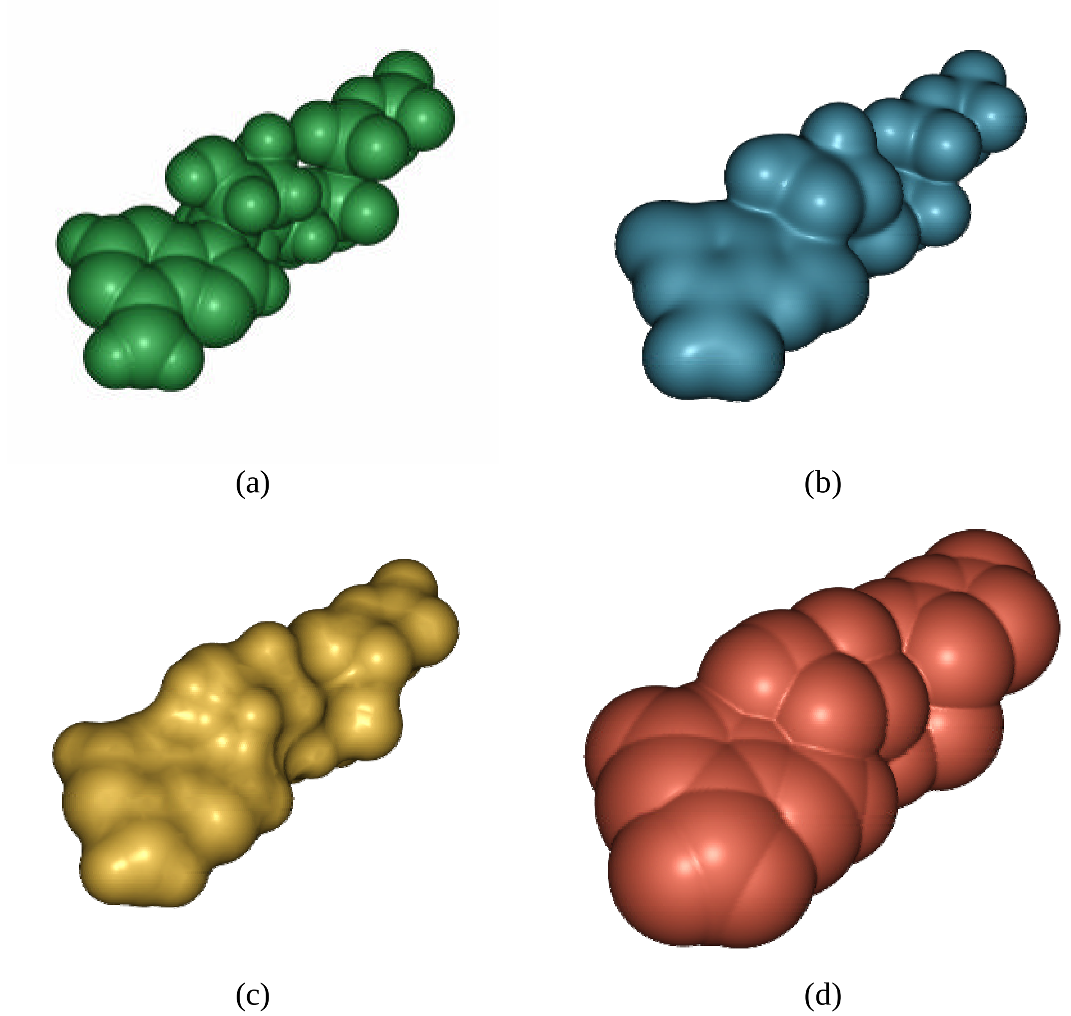
model it can be calculated quickly also for very large systems. It is smooth which makes it possible to calculate curvatures for every point on the surface and it comprises a model that provides more information than the simple 3D hard-sphere arrangement of the van der Waals Surface. Therefore the solvent 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]:

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

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:

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

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).

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

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 (a_i) at the point from 2π (eq. 2-15).

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

eq. 2-14

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

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

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*:

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

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*.

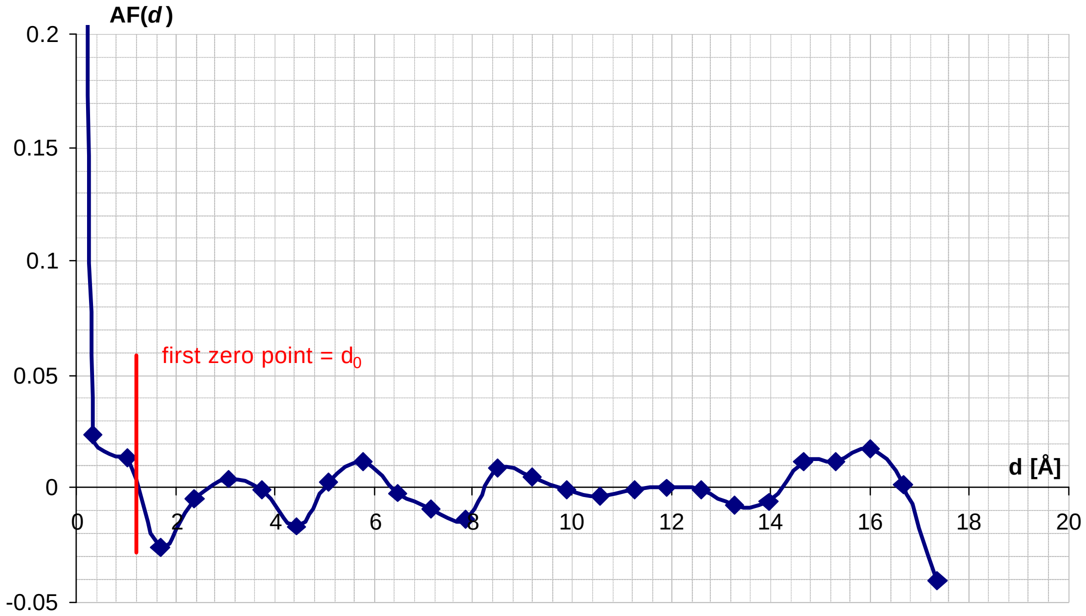


Figure 2-4: Surface autocorrelation function.

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 ith 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 ith 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].

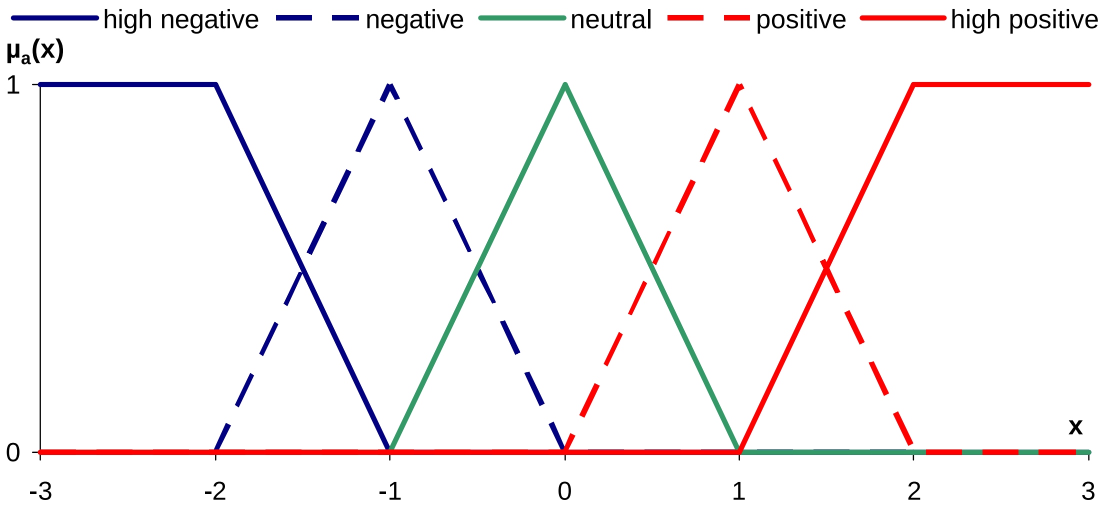


Figure 2-5: Shape of the membership functions in a linguistic variable.
The sets can be used to describe the electrostatic potential on a molecular surface. The variable x represents the autoscaled property values on the surface points and the five classes represent the highly minus, minus, neutral,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

$$q_i = q_{i-1} + \frac{a_i}{\sum_{b}^{border} a_b} \cdot 2p \quad \text{with } a_i = \angle(\mathbf{p}_i, \mathbf{p}_c, \mathbf{p}_{i-1}). \qquad \text{eq. 2-20}$$

where q_i and q_{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, a_i is the angle formed by these points and a_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(f), where f_i and f_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:

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

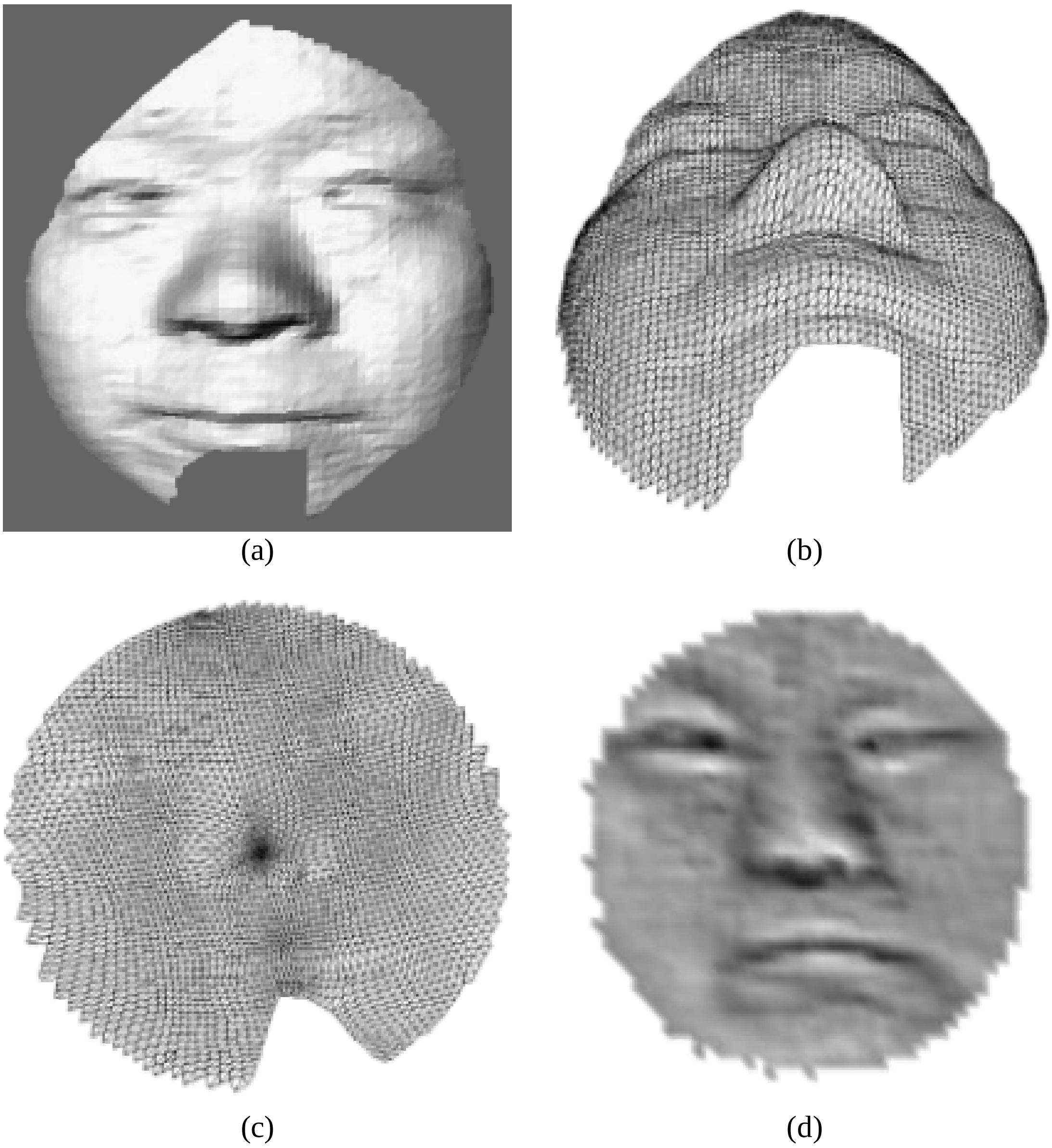
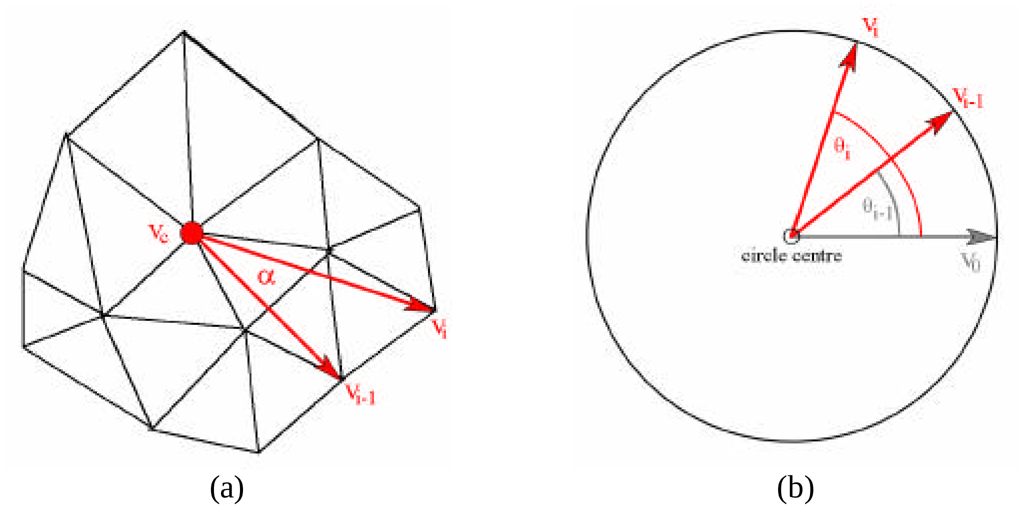


Figure 2-6: A surface patch, its harmonic map and harmonic shape image.
The pictures are redrawn from [144] and show a surface patch of a human face in shaded (a) and wireframe (b) representation. From that patch a harmonic map can be generated (c) which is subsequently resampled into a harmonic shape 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, a is the 3D angle and q_i and q_{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 f_i. The gradient of eq. 2-21 (eq. 2-22) yields eq. 2-23 as the components of a linear equation system.

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

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

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:

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

Taking a single row of the linear system, representing the energy function of a distinct mapping f_i eq. 2-24 is the first derivative of the energy function of f_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 f_i:

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

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 f_i^a, f_j^a and f_b^a leads to eq. 2-27, which is suitable for the matrix representation of the equation system (eq. 2-28).

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

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

Hence the systems of linear equations are

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

with

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

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

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

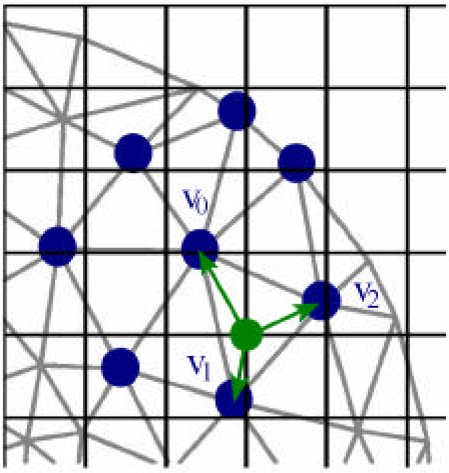


Figure 2-8: Interpolation scheme for the generation of harmonic shape images

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:

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

$$v_G = a \cdot v_0 + b \cdot v_1 + g \cdot v_2 \qquad \text{eq. 2-32}$$

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:

$$r = \frac{2g}{N}, \quad q = \frac{px}{4g} \text{ with } g = \max(|x|, |y|) \qquad \text{eq. 2-33}$$

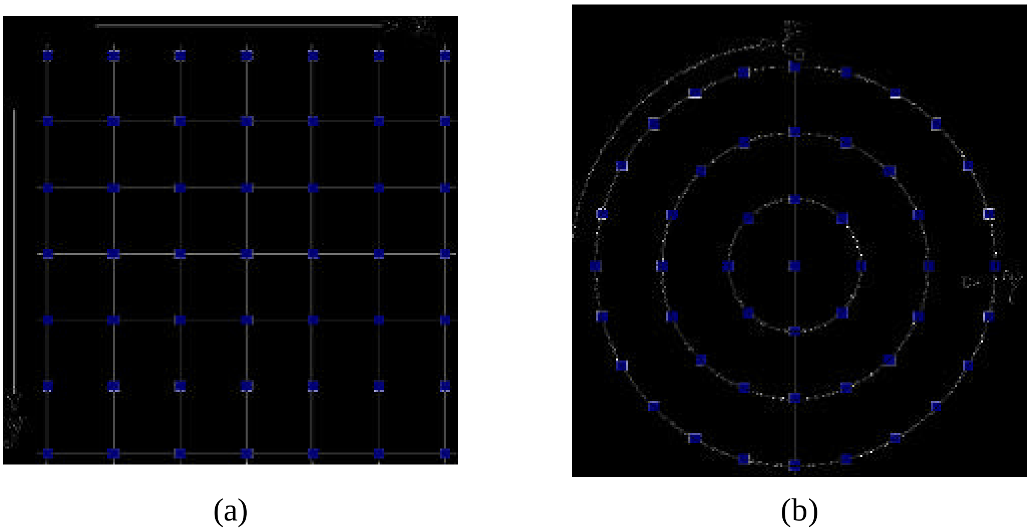
$$\begin{aligned} |x| = g: \qquad & x = 2\left(g - x)\frac{y}{|y|} + \frac{xy}{g}\right) \\ |y| = g: \qquad & x = 2y - \frac{xy}{g} \end{aligned} \qquad \text{eq. 2-34}$$

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 **p** and **q**:

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



(a) (b)

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 *q*, 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**:

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

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,

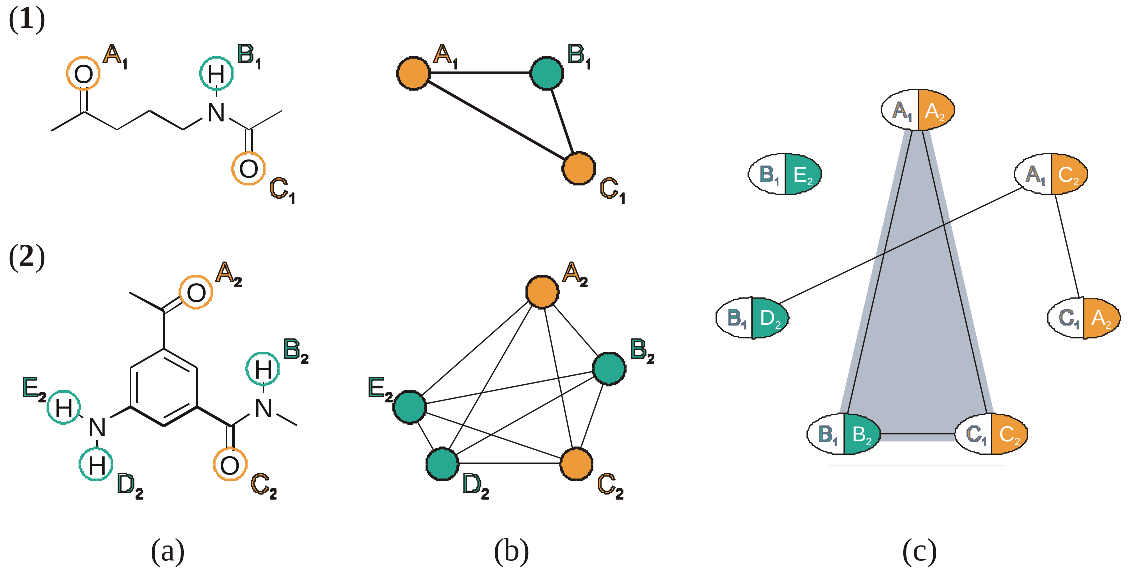


Figure 2-10: Pharmacophore match by maximum common subgraph isomorphism.
The pharmacophoric points on each molecule (a) are transformed into completely connected graphs (b). In the association graph (c) that is formed by the combination of all similar nodes in both graphs only those pairs are connected 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.