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

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.

1. INTRODUCTION

Noncovalent intermolecular interactions can be described in terms of complementary molecular surfaces coming into contact with each other. Comparison of molecular surfaces, based on their shape and/or physicochemical characteristics, can highlight and explain similarities in chemical and biological properties.[1−5] Analysis of local surface similarities may help interpret the biological activity of drug candidate molecules on a structural basis,[6] and surface complementarity is one of the most important heuristics built into docking algorithms.[7] It is therefore not surprising that a large number of papers have been published on molecular surface comparison methodologies.[2,8−16] By comparing the surface-mapped physicochemical properties of a set of ligands known to bind to the same receptor site, it should be possible to identify those features that play an important role in binding.

The presence of local surface similarities may not be immediately obvious from the chemical structures of the ligands and therefore their detection can contribute to a better understanding of the structural basis of biological activity. To facilitate this analysis we developed a method for detecting local surface similarities based on shape and surface-mapped molecular properties. Our approach is based on graph theory and a computer vision technique called Harmonic Shape Image Matching,[17] 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. Here we report the first implementation of the method and present the results from a series of tests on eight thermolysin inhibitors and four dihydrofolate reductase ligands. [Readers wishing to use the SURFCOMP package are kindly requested to contact the corresponding author directly.]

2. METHODS

Molecular surfaces are usually represented by triangle meshes containing up to several thousand points. It has been shown earlier that the problem of finding similarities between 3D point sets is equivalent to the maximum common subgraph or maximum subgraph isomorphism problem.[18,19] A widely used method in graph theory is that of Barrow and Burstall[20] which builds up an association graph followed by clique detection[21] to find the maximum common subgraphs between two query graphs. Unfortunately this is an NP-complete problem,[22] which makes it impossible to use the complete set of points of complex surface objects. 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. Cosgrove et al. reported such an application of subgraph isomorphism to molecular surface comparison.[6] They described the surfaces by patches of the same shape type and used local geometry parameters to decide which patches could overlap, but they did not consider the chemical environment of points on the surface.

Our approach is to generate a representation of the surfaces using slightly overlapping circular patches and keep track only of a set of shape critical points (*CP*) corresponding to the centers of those patches. The idea of critical points was explored by Connolly's docking algorithm[1] which was later improved by Lin et al.[3] 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 1

* Corresponding author phone: ++43 (1) 86634-452; fax: ++43 (1) 86634-727; e-mail: andras.aszodi@pharma.novartis.com.
† Novartis Institutes for BioMedical Research.
‡ Technische Universität Wien.

10.1021/ci0342371 CCC: $27.50
Published on Web 03/09/2004

Table 1. Complexity of the Association Graph at Different Steps of the Filtering Process Shown for the Comparison of 1THL (A) and 4TMN (B)

process step	section[a]	points A[b]	points B[b]	nodes	edges
at the beginning		1131	1265	1.47×10^6	2.04×10^{12}
after					
- critical point detection	2.1	27	29	553	274 841
- fuzzy property filter	2.3	24	27	162	17 982
- harmonic shape image filter	2.4	18	25	63	1260
- distance filter	2.5	18	25	63	359
- overlap filter	2.6	18	25	60	93

[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.

illustrates the complexity of the association graph at the initial stage and after every step of the algorithm.

For efficiency reasons we emphasize the simplification of the association graph which results in a set of smaller cliques that must be combined to reproduce the complete similarities between the molecular surfaces. We used a hierarchical clustering method to finally combine those cliques that represent the same geometrical transformation of one molecule onto the other.

The remaining part of this section describes in detail the selection of the critical points, the creation and filtering of the association graph, and the final clique detection and clustering (see also Figure 1).

2.1. Definition of Critical Points. From the complete set of points representing a molecular surface we extract a subset of shape critical points. To accurately describe the shape of the surface we used the first and second canonical curvature 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 cutoff range c_{CR}. This paraboloid is a parametrical approximation $S_p(u,v)$ 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 are obtained as the first and second eigenvalue of the Hessian matrix H, respectively,[23] (see eq 1)

$$H = \begin{bmatrix} \dfrac{\partial^2 S_p(u,v)}{\partial u^2} & \dfrac{\partial^2 S_p(u,v)}{\partial u \partial v} \\ \dfrac{\partial^2 S_p(u,v)}{\partial v \partial u} & \dfrac{\partial^2 S_p(u,v)}{\partial v^2} \end{bmatrix} \tag{1}$$

The signs of the canonical curvatures at each point are used to assign it to one of three shape classes: convex regions have two negative, concave two positive, and saddle shaped ones one positive and one negative curvature. Hence we define two classes of critical points: A *peak* is a convex point with minimal curvature and a *valley* is a concave point with maximal curvature in a certain neighborhood n_{CP}, corresponding to a "dip" or "hole" on the surface. To keep the initial set of critical points as small as possible we do not consider saddle points.

The *CP* algorithm investigates every convex or concave point on the surface and adds every *peak* or *valley* it finds to the set of critical points. Figure 2 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 in really concave but saddle shaped regions.

2.2. The Association Graph. The vertices of the association graph correspond to pairs of critical points, $pp_{ij} = (CP_{iA}, CP_{jB})$, from the two surfaces compared. 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. By definition, edges should be drawn between every two pairs of critical points that do not have a critical point in common (see Figure 1b), but for computational reasons no edges are considered before the application of the distance filter.

2.3. Fuzzy Property Filter. It is advisable to remove those critical point pairs from the association graph that do not have the same chemical environment. Each vertex is thus checked by a chemical filter to ensure that the corresponding critical points have similar chemical properties. We used fuzzy sets and linguistic variables[24] to express the similarity between chemical properties mapped onto the surface and applied a defuzzification function, D_{LV} (see eq 4), introduced by Exner et al.[16] as a similarity measure:

The chemical property mapped to a critical point is classified by a family of five fuzzy sets A_i (with overlapping membership functions $\mu_i(x)$) over the standardized range of property values X; see eq 2 and Figure 3) which are grouped into a linguistic variable LV (eq 3).

$$A_i = \{(x,\mu_i(x)) | x \in X\} \tag{2}$$

$$LV = \{A_1, A_2 \ldots A_5\} \tag{3}$$

Two critical points, x and y, are then compared by their linguistic variables LV_x and LV_y by the following measure

$$D_{LV}(x,y) = \sum_{i=1}^{5} \frac{w_i |\mu_i(x) - \mu_i(y)|}{w_i(\mu_i(x) + \mu_i(y))} \tag{4}$$

where μ_i and w_i be the ith membership function and its weight. The range of D_{LV} is between 0 and 1, with zero indicating identity and one complete dissimilarity. The different weights w_i are set to 1.

Based on this fuzzy dissimilarity function we can define a crisp filter condition that eliminates all pairs of critical points that are more dissimilar than a certain fuzzy threshold F.

2.4. Harmonic Shape Image Filter. The comparison of pairs of points alone is not sufficient to scan for surface similarities. It is also important to consider patches of the molecular surface around them and compare the shape of these patches with each other to establish whether both points are embedded in similar regions and how they are best

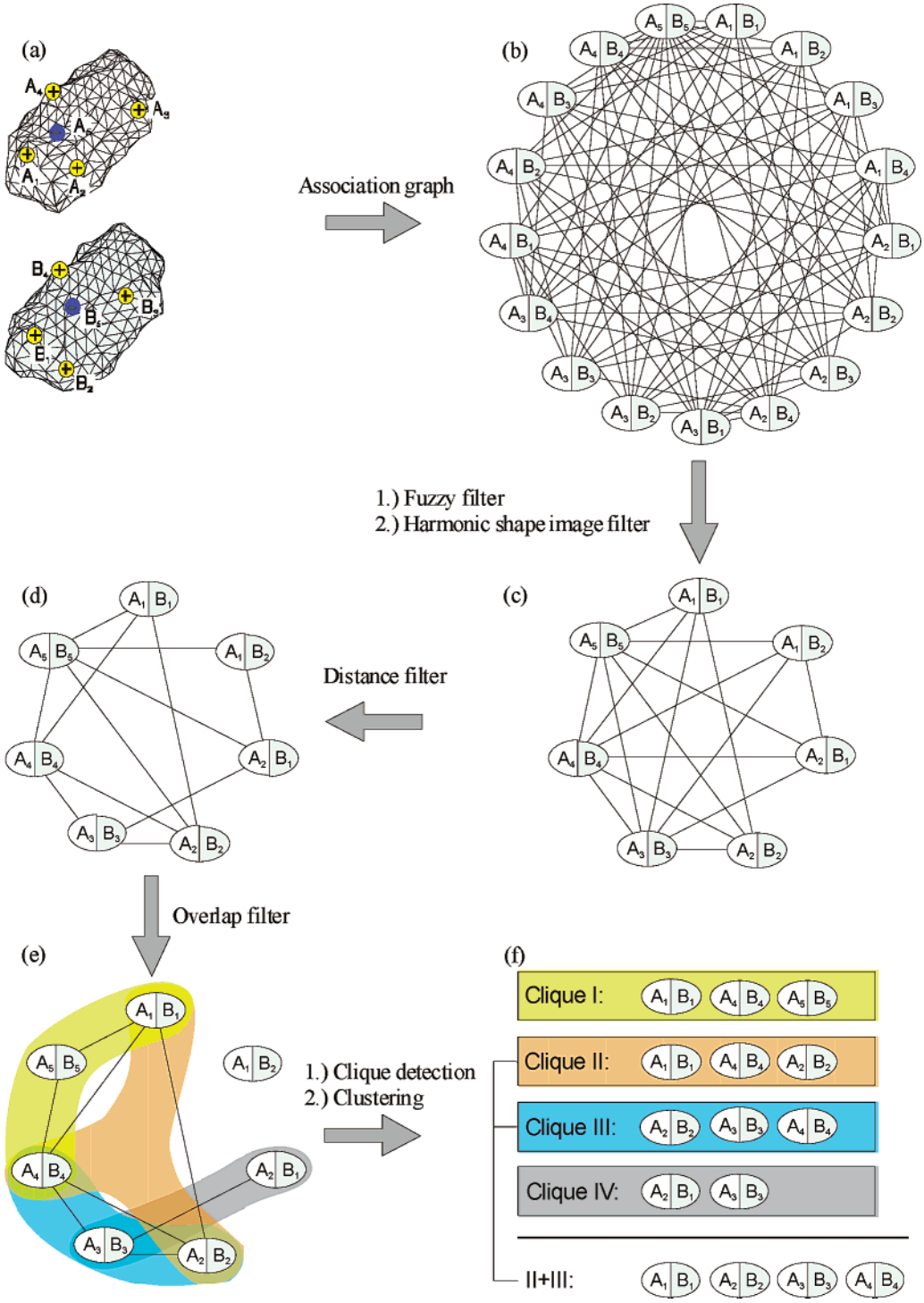


Figure 1. Distribution of peak (red) and valley (green) critical points over the surface of N-(1-(2(R, S)-carboxy-4-phenyl-butyl) cyclopentylcarbonyl)-(S)-tryptophan, a thermolysin inhibitor.

oriented relative to each other. Harmonic shape images[17] provide a methodology to compare patches and to define a relative orientation. They serve as 2D representations of 3D surface regions, and the comparison of 3D patches is thus reduced to a rather simple 2D image comparison.

The 2D images are generated via harmonic mapping[25] which consists of "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.,[26] this minimal distortion criterion can be formulated using a physical analogy. Let us imagine 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. Zhang reported a complete procedure for the harmonic mapping of circular surface patches of triangulated meshes.[27]

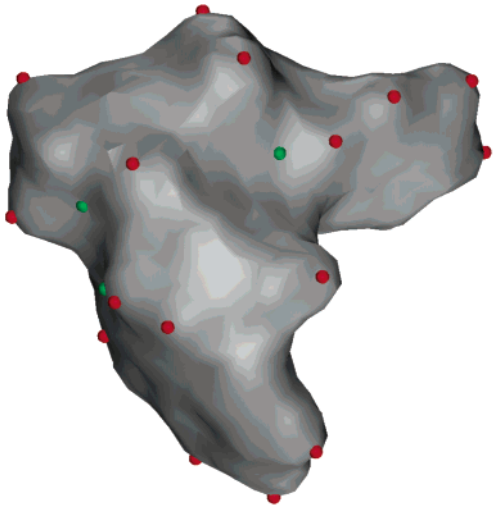


Figure 2. Overview of the similarity detection algorithm. Starting from two molecular surfaces the critical points are identified (a) and an initial association graph is built (b), which is then further simplified by the fuzzy and harmonic shape image filter (c), the distance filter (d), and the overlap filter (e). Of the final association graph the cliques are detected (green, orange, blue, and gray regions) and merged (clique II and II in this example) to yield the maximal surface similarity (f).

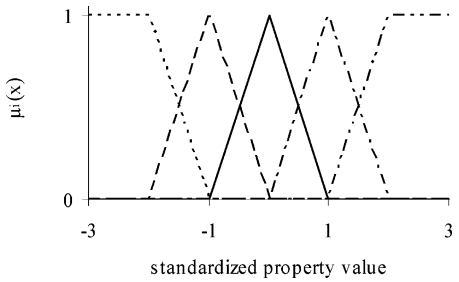


Figure 3. Shape of the membership functions of the five fuzzy sets in the linguistic variables that describe the chemical character of a critical point. E.g. in the case of the electrostatic potential, the functions correspond to highly negative, negative, neutral, positive, and highly positive areas of the surface proceeding from left to right.

It can be shown[25] 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 that 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. While in principle any scalar function or property defined on a molecular surface can be associated with the harmonic image vertices, in this work we focused on geometrical descriptors.

Having mapped a pair of 3D patches onto the unit disk in 2D, the comparison consists of rotating the images relative to each other until the orientation with maximum shape similarity is found. Shape is considered to be a curvature value, calculated for every vertex in the patch (see below).

As the vertex topology of the harmonic images is almost always different, the comparison must be based on a regular grid scheme that is identical for all patches. Zhang[27] resamples the harmonic maps 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. This approach 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.[28] A circular grid also has a higher symmetry than a rectangular grid which allows a faster computation of the relative rotations.

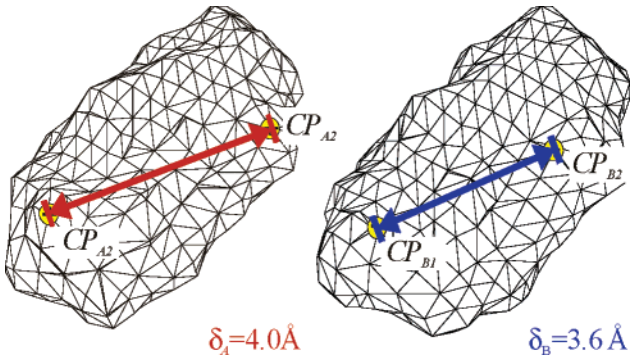


Figure 4. Distance filter, illustrating how δ_A (left) and δ_B (right) are measured and compared.

The similarity of two harmonic images can be expressed by the normalized correlation coefficient R between the vectors p and q representing the sequence of corresponding grid points. The correlation coefficient is a function of the rotation angle θ, and the similarity is defined as the maximum of this function.

The harmonic image filter further reduces the number of vertices in the association graph by eliminating all those pairs that have a correlation coefficient below a specific shape threshold R. Furthermore, due to the one-to-one correspondence, a set of corresponding point pairs around the critical points CP_A, CP_B can be established which define the best overlap of the surrounding patches. This best overlap position is used later to check the simultaneous overlap of two point-pairs and to construct a rigid body transformation between the detected similar regions.

2.5. Distance Filter. The fuzzy and harmonic image filters consider only single pairs of critical points (*pp*), but 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 so that those which can overlap at the same time are connected.

A simple but effective criterion is the difference of the distances of two point pairs on 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 cp_{A1}, cp_{B1} and cp_{A2}, cp_{B2}, the distances δ_A and δ_B are

$$\delta_A = ||cp_{A1} - cp_{A2}||(A)$$

$$\delta_B = ||cp_{B1} - cp_{B2}||(B) \qquad (5)$$

the Euclidean distances between the two critical points on surfaces A and B (see also Figure 4).

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 minimal distance δ_{min}. The minimal distance is introduced to avoid connections between very close critical point pairs which represent essentially the same regions.

2.6. Overlap Filter. The distance filter checks if two pairs are at an appropriate distance for simultaneous overlap, but

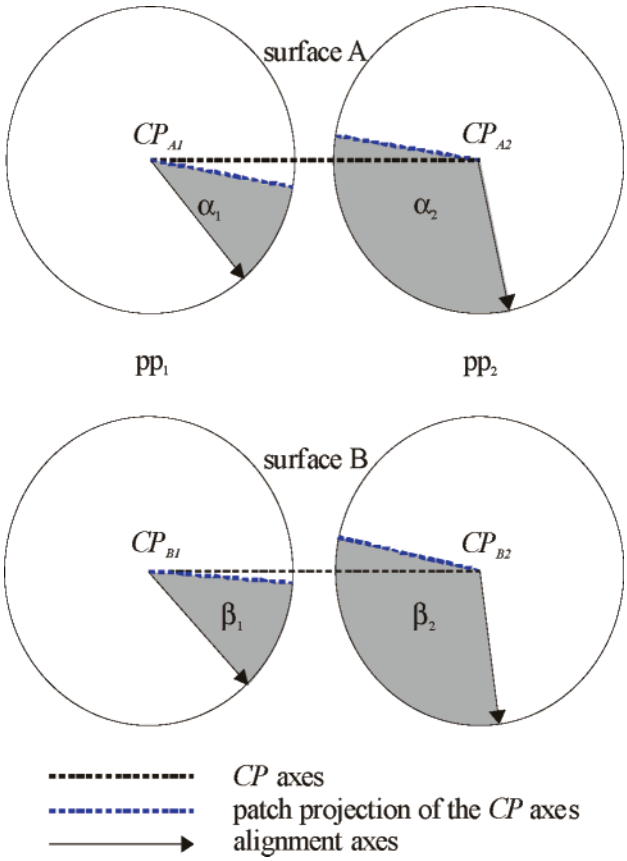


Figure 5. Illustration of the overlap filter. The axes between the two patches on both surfaces (black stippled lines) are projected onto the harmonic map of the surface patch, and the angles between that projections and the axes that define "north" (0°) in the optimal alignment of the patchpairs pp_1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

harmonic image matching provides us with 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 orientation of the connecting axes on surface A and B. In Figure 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 optimal 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 bearing from one critical point patch to the other.

The filter computes the heading differences ϕ_1, ϕ_2 for both *CP* patch pairs (eq 6) and removes the connection between them, if none of them is within a certain angular tolerance ϕ_{tol}.

$$\phi_1 = |\beta_1 - \alpha_1|$$
$$\phi_2 = |\beta_2 - \alpha_2| \qquad (6)$$

2.7. 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. We used the algorithm of Bron and Kerbosch[21] 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. These small primary cliques can be combined into larger clusters that represent different sets of corresponding points on both surfaces.

For each cluster we can generate a rigid body transformation based on all correspondences detected by the harmonic shape image matches for the patches around the critical points. The transformations have been calculated by a least-squares fit[29] 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. From the large set of initial small clusters, those with high RMSD values are eliminated (above 2.0 Å), and the remaining clusters are subjected 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.

Beside single linkage we also examined complete and average linkage but could not find any differences in the quality of the results. Because of that and because of the fact that single-linkage can be implemented more efficiently than complete and average linkage, we used single linkage in all our experiments.

2.8. Molecular Surfaces and Properties. There are several ways to define molecular surfaces. Among the most often used molecular surfaces in computational chemistry are the solvent accessible surfaces, which were first introduced by Lee and Richards[30] and popularized by Michael Connolly's MS program.[31] We generated the surfaces of the test molecules with the MOLCAD module[32] of Sybyl 6.9[33] using a probe sphere radius of 1.4 Å and a point density of 3 dots per Å^2.

The two canonical curvatures (see eq 1) were appropriate for the critical point detection, but for the use in the harmonic shape image filter we needed a univariate representation of the local curvature. We made use of the surface topography index (STI, eq 7) as implemented in MOLCAD[34] that assigns a real-valued curvature descriptor to each point according to its first and second canonical curvature (cc_1, cc_2)

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

$$\mathrm{STI} = \frac{cc_1 + 3 \cdot cc_2}{cc_1} \text{ if } cc_1 \le 0 \text{ and } cc_2 < 0 \text{ or if } (cc_1 > 0 \text{ and } cc_2 \le 0) \text{ and } |cc_1| \le |cc_2| \qquad (7)$$

We mapped two physicochemical properties onto the molecular surfaces: the electrostatic potential (ESP) and the lipophilic potential (LP). The ESP can be calculated by

Coulomb's law if net atomic charges, q_i, are available (eq 8, with r_i and r_j denoting the position of the surface dots and atoms, respectively). We used atomic point charges that reproduced the electrostatic potential as calculated by the semiempirical program MOPAC.[35]

$$\mathrm{ESP}(v_i) = \sum_{j=1}^{N} \frac{q_j}{||r_i - r_j||} \tag{8}$$

The hydrophobic effect plays an important role in drug−receptor interactions. While not a molecular property itself, it can be described empirically by, for example, the *n*-octanol/water partition coefficient (log*P*). Ghose and Crippen[36] assembled a table of fragmental log*P* 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", $\mathrm{LP_{HM}}(v_i)$, to every point v_i on the surface similar to the ESP[37]

$$\mathrm{LP_{HM}}(v_i) = \frac{\sum_{j}^{N} f_j \cdot g(d_{jj})}{\sum_{j}^{N} g(d_{ij})} \text{ with } g(d_{ij}) = \frac{e^{-C_1 C_1} + 1}{e^{C_1}(d_{ij} - C_2) + 1} \tag{9}$$

where d_{ij} is the distance between the surface point i and the atom j, and C_1 and C_2 are experimental constants.

3. RESULTS AND DISCUSSION

We assembled two test sets of ligand structures: thermolysin inhibitors and dihydrofolate reductase (DHFR) inhibitors together with folic acid. The thermolysin set was subject to an earlier surface similarity search performed by Cosgrove et al.[6] with their SPAt program. The DHFR compounds were assembled from structures published by Li et al.[38] (methotrexate, trimetoprim, and Br-WR92210) and Davies et al.[39] (folic acid).

All the structures were extracted from crystallographic data of protein/ligand complexes available in the Brookhaven Protein Data Bank (PDB).[40] To compare the overlays generated by our method with the experimental alignments of the different ligands in the proteins' active sites, we superimposed the complexes in the PDB 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.[33] For each structure the molecular surfaces and properties were calculated as described above using the experimental parameters as summarized in Table 2.

We performed exhaustive comparisons with all pairs of structures in the single data sets and manually inspected the overlays found by the method to ensure that the results really represent the expected molecular alignments. A single comparison took about 75 s (± 15 s) on a 2.4 GHz Intel Xeon processor with 2GByte of RAM, running under Linux (kernel version 2.4.19).

Table 2. Experimental Conditions Used in the Thermolysin and DHFR Experiments

filter parameter	symbol	section[a]	value[b]	property[c]
curvature cutoff range	c_{CR}	2.1	2.0 Å	
neighborhood	n_{CP}	2.1	2.0 Å (thermolysin) 1.0 Å(DHFR)	
fuzzy threshold	F	2.3	0.3	ESP or LP
shape threshold	R	2.4	0.6	STI
distance tolerance	t	2.5	1.0 Å (thermolysin) 2.0 Å (DHFR)	
minimum distance	δ_{min}	2.5	0.5 Å	
angular tolerance	ϕ_{tol}	2.6	15.0°	

[a] The section in the text where the filter is described. [b] If different values were chosen for the thermolysin and dihydrofolate reductase (DHFR) data set, it is noted in parentheses. [c] The molecular-surface property applied to the specific filter (see also section 2.8): electrostatic potential (ESP), lipophilic potential (LP), shape topology index (STI).

Table 3. Overlays of Thermolysin Inhibitors Performed with Electrostatic Potential (ESP) and Lipophilic Potential (LP)

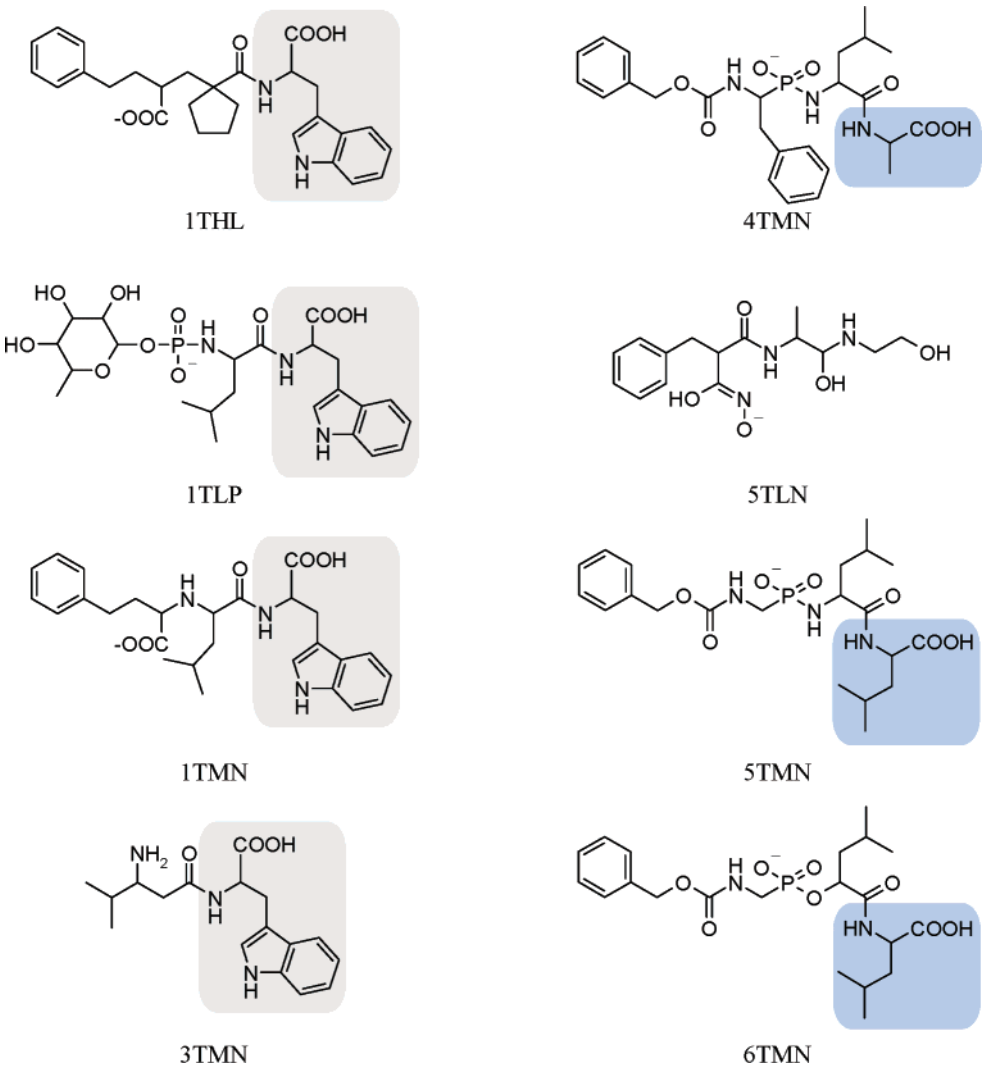
molecules		ESP		RMSD[a] [Å]		LP		RMSD[a] [Å]	
A	B	CPs[b]	points[c]	surf.	struct	CPs[b]	points[c]	surf.	struct
1THL	1TLP	8	553	1.15	0.58	6	441	1.60	1.04
	1TMN	12	711	1.77	0.40	8	554	1.55	0.31
	3TMN	7	366	1.04	0.33	5	368	1.12	0.55
	4TMN	5	431	1.07	1.18	4	349	0.98	0.95
	5TLN	4	227	1.93	5.68	2	181	1.78	5.11
	5TMN	4	336	1.04	1.20	2	169	0.98	7.08
	6TMN	7	439	1.00	0.63	2	228	0.89	0.73
1TLP	1TMN	10	630	1.73	0.53	3	309	1.35	1.39
	3TMN	7	471	1.26	0.46	5	424	1.52	1.20
	4TMN	7	446	2.16	1.29	2	188	1.51	6.01
	5TLN	5	342	2.13	7.00	4	335	2.50	1.27
	5TMN	7	454	0.93	0.63	3	165	0.63	1.22
	6TMN	7	409	1.12	0.59	5	282	0.79	1.04
1TMN	3TMN	2	193	0.93	1.00	7	393	2.50	0.75
	4TMN	7	446	1.05	1.03	6	417	1.44	0.80
	5TLN	2	145	0.99	5.14	2	205	1.61	6.25
	5TMN	9	464	1.21	0.93	2	222	0.71	0.84
	6TMN	11	610	1.26	0.99	5	426	1.49	0.86
3TMN	4TMN	3	255	1.36	1.42	5	339	2.07	5.45
	5TLN	3	252	1.99	2.90	2	116	0.58	6.91
	5TMN	3	254	1.18	1.51	5	363	1.68	1.18
	6TMN	2	180	1.26	4.28	3	283	1.39	0.67
4TMN	5TLN	5	383	3.52	5.83	2	169	1.17	6.22
	5TMN	5	320	0.75	0.43	2	168	0.52	0.54
	6TMN	6	409	0.83	0.58	4	312	1.90	0.49
5TLN	5TMN	2	175	1.34	2.31	2	176	1.44	3.37
	6TMN	2	153	1.77	5.78	2	188	1.55	1.18
5TMN	6TMN	20	975	0.51	0.08	19	965	0.55	0.05

[a] Root mean square deviation. [b] The number of critical points that build the cluster for that overlay. [c] 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).

3.1. Thermolysin Inhibitors. The structures of the eight thermolysin inhibitors in Chart 1 were extracted from the PDB. All molecules except 3TMN and 5TLN are complexed via a negatively charged carboxyl- or phosphate group to a zinc ion in the active site of the protein. Thus we placed single negative formal charges at these positions. 5TLN is also complexed to the zinc ion but via a hydroxamic acid group which is also charged. 3TMN does not show any complex binding to the ion at all and was left uncharged.

We performed two different experiments, one with the electrostatic and one with the lipophilic potential mapped

Chart 1. 2D Structures of Eight Thermolysin Inhibitors[a]



[a] The structures are identified by the PDB entry name of the corresponding protein/ligand complex.

onto the molecular surfaces (Table 3). Using the ESP we 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 (grey boxes) and structures with an aliphatic (alanine, leucine; blue boxes) residue at the C-terminal end. The tryptophan structures could be overlaid with a 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. 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.

The overlays found by the surface matching conducted with the LP as the chemical filter were in general not as good as the ESP results. 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 3, and example alignments are displayed in Figures 6 and 7. Our results agree with the alignments published earlier by Cosgrove et al.[6] for the same data set.

3.2. DHFR Ligands. The set of four dihydrofolate reductase ligands, the substrate folic acid and three inhibitors,

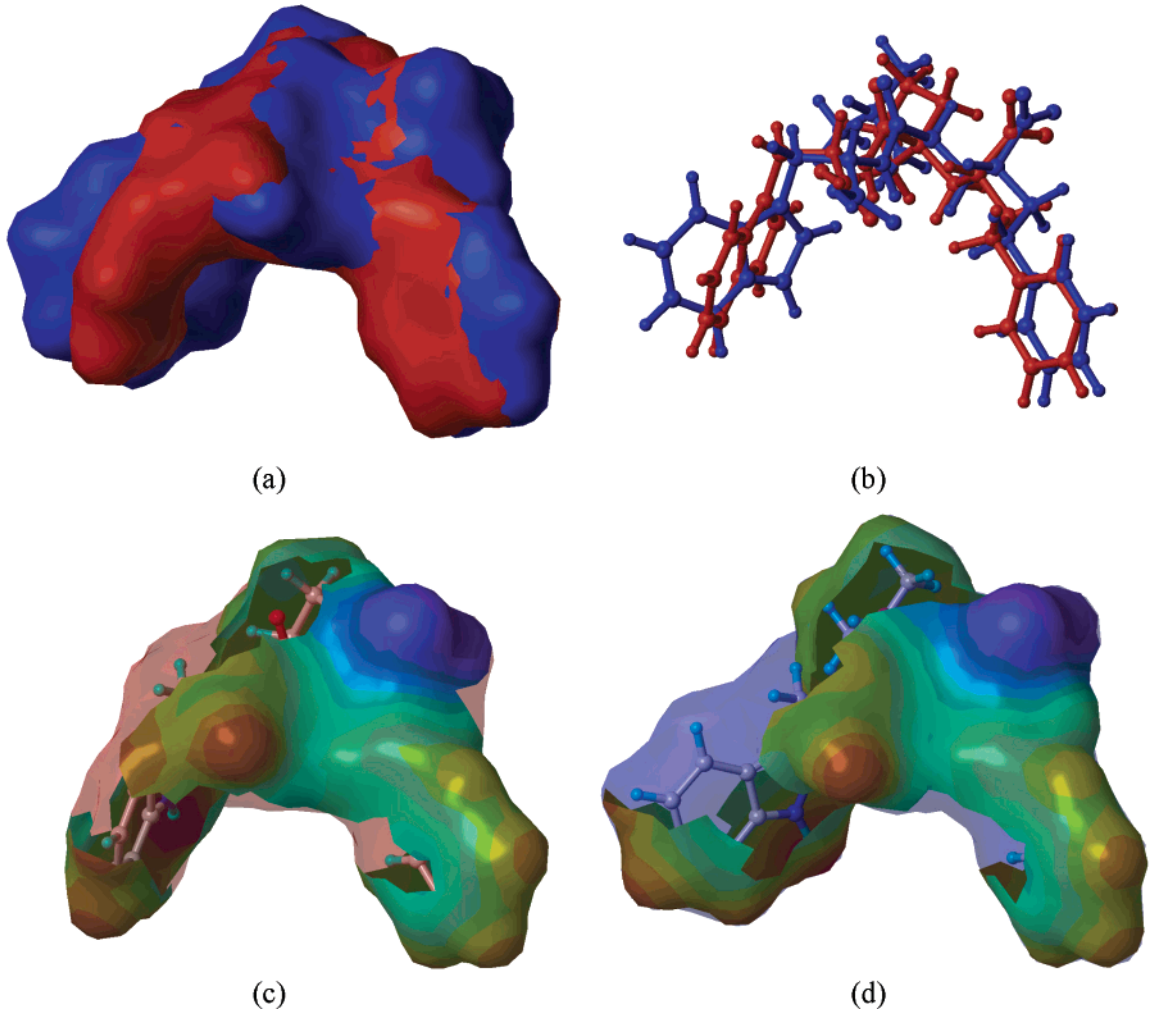


Figure 6. Surface alignment of 1THL (red) and 1TMN (blue). (a) and (b) display the alignment of the molecular surfaces and structures respectively based on the detected surface similarity. (c) and (d) show the similar surface regions of 1THL and 1TMN color coded by the electrostatic potential to illustrate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

was prepared from the PDB exactly as the thermolysin data set above. The common feature of all four structures is a nitrogen-containing heterocycle (pyrimidine or pteridine) 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 (see Chart 2).

We found that FOL could be aligned properly with MTX (Figure 8) and WRB especially around the heterocycles, but that the alignment with TMP is poorer although the amino groups at the heterocycles were aligned correctly. Most of the other alignments were not as good as expected. This is probably due to the fact that around the active parts the surface of the molecules are rather featureless and flat, thus only a few critical points are positioned at the substituents of the heterocycles. In the alignments these points are mixed with locally strong alignments on the side chains, leading eventually to incorrect overlaps. Another problem with these molecules is a local surface symmetry around the CPs on the amino-groups which leads to alignments that are locally correct but do not reproduce the actual relative positions at the binding sites. (e.g. MTX, WRB). The results are summarized in Table 4.

Table 4. Overlays of DHFR Ligands Performed under Electrostatic Potential (ESP) Conditions

molecules				RMSD[a] [Å]	
A	B	CPs[b]	points[c]	surf.	struct
FOL	MTX	6	449	1.36	1.23
	TMP	3	215	1.32	1.90
	WRB	4	257	0.99	1.26
MTX	TMP	4	216	0.64	1.63
	WRB	3	181	1.11	5.82
TMP	WRB	5	312	1.28	1.74

[a] Root mean square deviation. [b] The number of critical points that build the cluster for that overlay. [c] 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).

4. CONCLUSION

We demonstrated that our method is capable of detecting regions of local similarities between two molecular surfaces in a reasonable amount of time. The surface point correspondences can then be used to calculate superimpositions based on *partial* rather than global surface similarities. Since in most ligand−receptor interactions only a certain part of the ligand or the receptor is involved in the binding, our

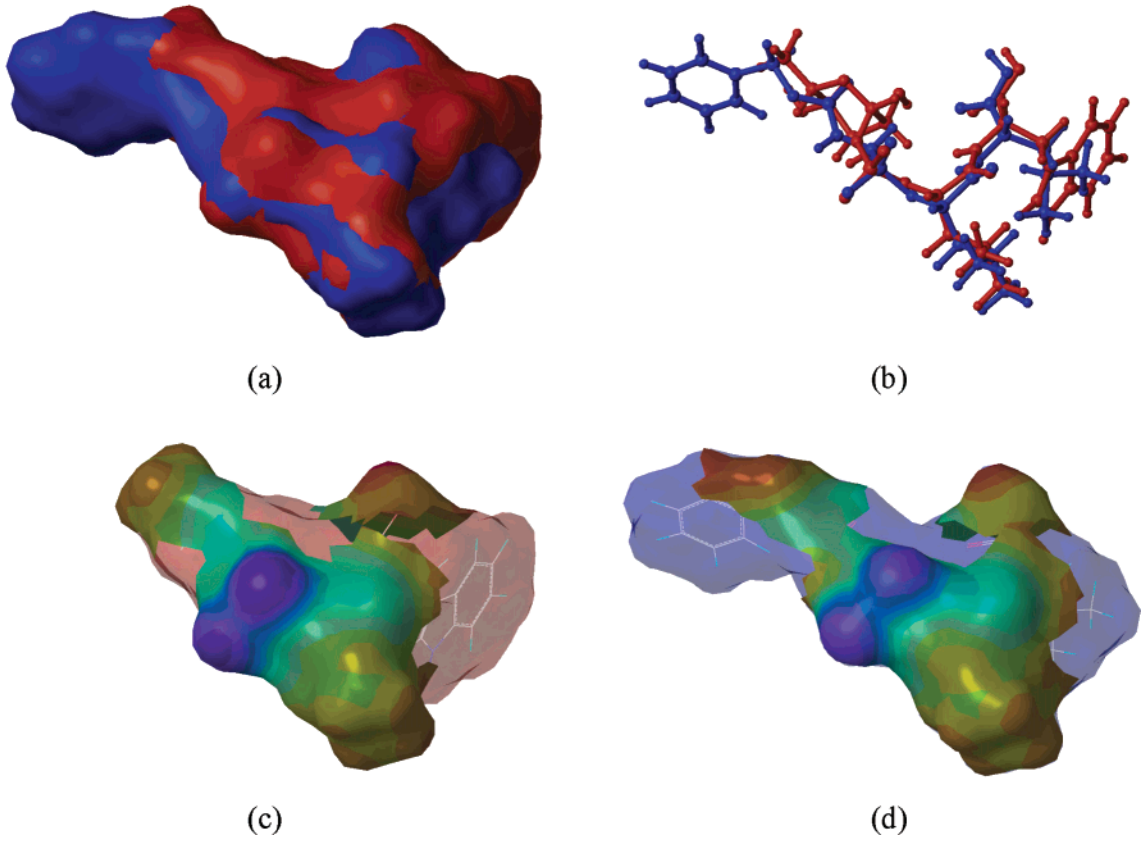
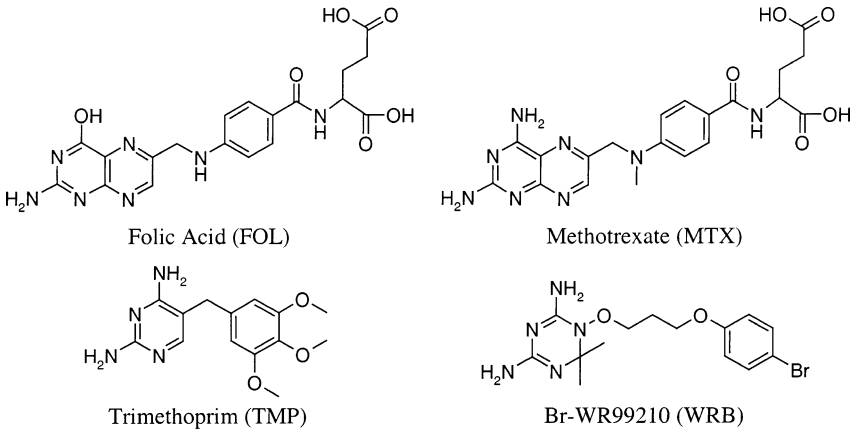


Figure 7. Surface alignment of 1TLP (red) and 6TMN (blue). (a) and (b) display the alignment of the molecular surfaces and structures respectively based on the detected surface similarity. (c) and (d) show the similar surface regions of 1TLP and 6TMN color coded by the electrostatic potential to illustrate their size and physicochemical similarity.

Chart 2. 2D Structures of Four DHFR Ligands



method can provide important insights into the mechanism of receptor binding even if the structure of the binding site is unknown.

The representation of the results as a hierarchical cluster of subalignments allows a deep insight into the nature of the similarity between the two molecules. On the way from the largest down to the smaller alignments one can easily identify the similar surface regions that reveal a good picture of the important stereochemical surface patterns (Figures 6−8).

With our approach it is possible to compare two surfaces using their shape and physicochemical properties at the same time. The results therefore represent a simultaneous match of geometry and chemistry which gives a deeper insight into the molecular analogies than shape alone. The filter based procedure also provides a very flexible framework that can be adapted to a large variety of surface similarity problems.

Although the results presented here were obtained using only ESP and LP mapped onto the molecular surfaces, it is straightforward to use other relevant physicochemical properties such as hydrogen bonding donor/acceptor parameters within the same framework. This could be implemented either for the whole surface as described by Exner et al.[16] or for each single site as proposed by Raevsky et al.[41,42]

Our method is currently applicable to rigid molecular conformations only. However, as the comparison runs reasonably fast, it is possible to combine it with a conformational analysis and scan a set of low-energy conformations

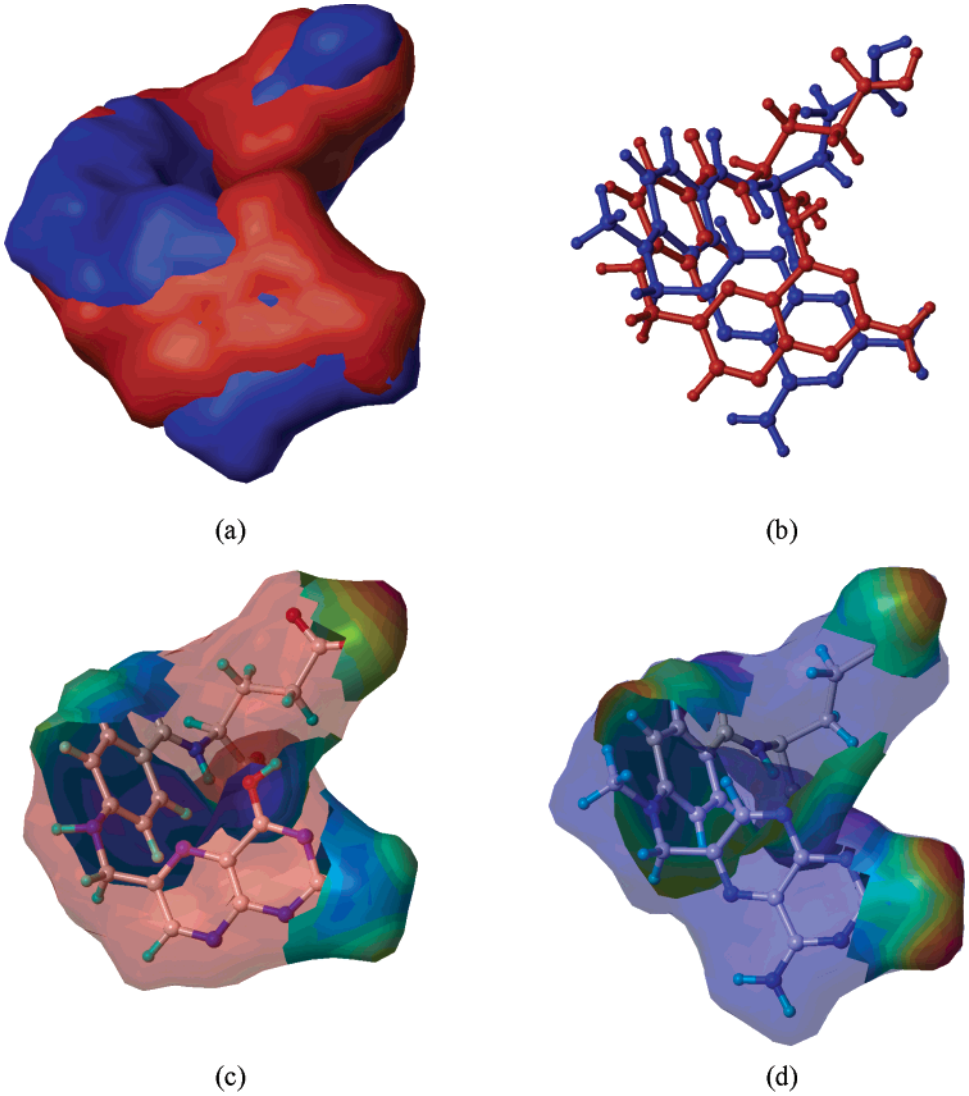


Figure 8. Surface alignment of FOL (red) and MTX (blue). (a) and (b) display the alignment of the molecular surfaces and structures respectively based on the detected surface similarity. (c) and (d) show the similar surface regions of FOL and MTX color coded by the electrostatic potential to illustrate their size and relative physicochemical similarity.

of each molecule as expected from a complete 3D molecular similarity analysis.

In addition to the investigation of the relative configurations of ligands within a specific binding site, presented in this publication, the method may be useful in the context of other molecular modeling problems. With only minor adaptations it will be possible to compare the surface of proteins or parts of proteins with each other on the same time scale. This could help to reveal the structural similarities between functionally related proteins that do not show significant structural similarity.

Together with conformational analysis, the method could be applied to predict a common binding mode in an unknown receptor if a set of active compounds is known. This binding mode could be represented as the largest common set of similar patches on all surfaces of the data set. From such a set it should be straightforward to define a surface model that could be used to search structural databases for similar compounds or serve as input to a QSAR methodology. In combination with efficient scoring functions our method could thus be used to search a large set of molecules for similarities that are not discovered by conventional structure similarity algorithms.

ACKNOWLEDGMENT

We would like to thank Drs Adrienne James and Torsten Schindler for their helpful comments and suggestions.

REFERENCES AND NOTES

(1) Connolly, M. L. Shape complementarity at the hemoglobin $\alpha\beta$ subunit interface. *Biopolymers* **1986**, *25*, 1229−1247.
(2) Fischer, D.; Norel, R.; Wolfson, H.; Nussinov, R. Surface motifs by a computer vision technique: Searches, detection, and implications for protein−ligand recognition. *Proteins: Struct., Funct., Genet.* **1993**, *16*, 278−292.
(3) Lin, S. L.; Nussinov, R.; Fischer, D.; Wolfson, H. J. Molecular surface representations by sparse critical points. *Proteins: Struct., Funct., Genet.* **1994**, *18*, 94−101.
(4) 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.* **1995**, *252*, 263−273.
(5) 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.* **2000**, *57*, 1970−1977.

(6) Cosgrove, D.; Bayada, D.; Johnson, A. A novel method of aligning molecules by local surface shape similarity. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 573−591.
(7) Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. Principles of docking: an overview of search algorithms and a guide to scoring functions. *Proteins: Struct., Funct., Genet.* **2002**, *47*, 409−443.
(8) Bladon, P. A rapid method for comparing and matching the spherical parameter surfaces of molecules and other irregular objects. *J. Mol. Graphics* **1989**, *7*, 130−137.
(9) Masek, B. B.; Merchant, A.; Matthew, J. B. Molecular skins: A new concept for quantitative shape matching of a protein with its small molecule mimics. *Proteins: Struct., Funct., Genet.* **1993**, *17*, 193−202.
(10) Perkins, T. D. J.; Mills, J. E. J.; Dean, P. M. Molecular surface-volume and property matching to superpose flexible dissimilar molecules. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 479−490.
(11) Rinaldis, M. D.; Ausiello, G.; Helmer-Citterich, G. C. M. Three-dimensional profiles: a new tool to identify protein surface similarities. *J. Mol. Biol.* **1998**, *284*, 1211−1221.
(12) Ritchie, D. W. A. Protein docking using spherical polar Fourier correlations. *Proteins: Struct., Funct., Genet.* **2000**, *39*, 178−194.
(13) Poirrette, A. R.; Artymiuk, E. J.; Rice, D. W.; Willett, P. Comparison of protein surfaces using a genetic algorithm. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 557−569.
(14) 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.* **1998**, *17*, 27−36.
(15) Pickering, S. J.; Bulpitt, A. J.; Efford, N.; Gold, N. D.; Westhead, D. R. AI-based algorithms for protein surface comparisons. *Comput. Chem.* **2001**, *26*, 79−84.
(16) Exner, T. E.; Keil, M.; Brickmann, J. Pattern recognition strategies for molecular surfaces. I. Pattern generation using fuzzy set theory. *J. Comput. Chem.* **2002**, *23*, 1176−1187.
(17) Zhang, D.; Herbert, M. Harmonic Maps and their applications in surface matching. *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR '99*) **1999**.
(18) 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.* **1993**, *7*, 83−102.
(19) Brint, A. T.; Willett, P. Algorithms for the identification of three-dimensional maximal common substructures. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 152−158.
(20) Barrow, H. G.; Burstall, R. M. Subgraph isomorphism, matching relational structures and maximal cliques. *Inf. Process. Lett.* **1976**, *4*, 83−84.
(21) Bron, C.; Kerbosch, J. Algorithm 457 − Finding all cliques of an undirected graph. *Commun. ACM* **1973**, *16*, 575−577.
(22) 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.
(23) Zachmann, C. D.; Heiden, W.; Schlenkrich, M.; Brickmann, J. Topological analysis of complex molecular surfaces. *J. Comput. Chem.* **1992**, *13*, 76−84.
(24) Zadeh, L. A. Fuzzy Sets. *Inform. Control.* **1965**, *8*, 338−353.
(25) Eells, J.; Sampson, L. Harmonic mappings of Riemannian manifolds. *Am. J. Math.* **1964**, *86*, 109−160.
(26) Eck, M.; DeRose, T.; Duchamp, T.; Hoppe, H.; Lounsbery, M.; Stuetzle, W. *Multiresolution analysis of arbitrary meshes*; University of Washington: Seattle, 1995.
(27) Zhang, D. Harmonic Shape Images: A 3D free-form surface representation and its applications in surface matching. Ph.D. Thesis, Carnegie Mellon University, U.S.A, 1999.
(28) Mukundan, R.; Ramakrishnan, K. R. *Moment functions in image analysis*; World Scientific: Singapore, 1998.
(29) McLachlan, A. Gene duplications in the structural evolution of chymotrypsin. *J. Mol. Biol.* **1979**, *128*, 49−79.
(30) Lee, B.; Richards, F. M. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **1971**, *55*, 379−400.
(31) Connolly, M. L. Solvent-accessible surfaces of proteins and nucleic acids. *Science* **1983**, *221*, 709−713.
(32) Brickmann, J.; Goetze, T.; Heiden, W.; Moeckel, G.; Reiling, S.; Vollhardt, H.; Zachmann, C. D. Interactive visualization of molecular scenarios with MOLCAD/SYBYL. *Data Visualization Mol. Sci.* **1995**.
(33) *SYBYL 6.9*; Tripos Inc.: St. Louis, MO, 2003.
(34) Heiden, W.; Brickmann, J. Segmentation of protein surfaces using fuzzy logic. *J. Mol. Graphics* **1994**, *12*, 106−115.
(35) 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.* **1985**, *107*, 3902−3909.
(36) 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. Comput. Chem.* **1986**, *7*, 565−577.
(37) 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.* **1993**, *7*, 503−514.
(38) 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.* **2000**, *295*, 307−323.
(39) 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* **1990**, *29*, 9467−9479.
(40) 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.* **2000**, *28*, 235−242.
(41) 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.* **1992**, *11*, 49−63.
(42) Raevsky, O. A.; Skvortsov, V. S. 3D hydrogen bond thermodynamics (HYBOT) potentials in molecular modelling. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 1−10.1

CI0342371