

A New Method to Detect Related Function Among Proteins Independent of Sequence and Fold Homology

Stefan Schmitt^{1,2}, Daniel Kuhn¹ and Gerhard Klebe^{1*}

¹*Inst. of Pharmaceutical Chemistry, Univ. of Marburg Marbacher Weg 6, D-35032 Marburg, Germany*

²*Structural Chemistry Laboratory, AstraZeneca R&D Mölndal, S-43183 Mölndal, Sweden*

A new method has been developed to detect functional relationships among proteins independent of a given sequence or fold homology. It is based on the idea that protein function is intimately related to the recognition and subsequent response to the binding of a substrate or an endogenous ligand in a well-characterized binding pocket. Thus, recognition of similar ligands, supposedly linked to similar function, requires conserved recognition features exposed in terms of common physicochemical interaction properties *via* the functional groups of the residues flanking a particular binding cavity. Following a technique commonly used in the comparison of small molecule ligands, generic pseudocenters coding for possible interaction properties were assigned for a large sample set of cavities extracted from the entire PDB and stored in the database Cavbase. Using a particular query cavity a series of related cavities of decreasing similarity is detected based on a clique detection algorithm. The detected similarity is ranked according to property-based surface patches shared in common by the different clique solutions. The approach either retrieves protein cavities accommodating the same (e.g. co-factors) or closely related ligands or it extracts proteins exhibiting similar function in terms of a related catalytic mechanism. Finally the new method has strong potential to suggest alternative molecular skeletons in *de novo* design. The retrieval of molecular building blocks accommodated in a particular sub-pocket that shares similarity with the pocket in a protein studied by drug design can inspire the discovery of novel ligands.

© 2002 Elsevier Science Ltd. All rights reserved

Keywords: functional comparison of proteins; cavity comparison; data mining; *de novo* design; physicochemical properties

*Corresponding author

Introduction

Genomic sciences provide us with the sequences of the entire human genome and other important microbial pathogens.^{1–4} By means of proteomics and powerful bioinformatic tools^{5–9} it is hoped to determine gene variants that contribute to various multifactorial diseases or to detect genes that exist in certain infectious agents but not in humans. As a consequence, a large number of new suitable targets for drug intervention may be discovered. Presently structural genomics embark on high-throughput X-ray crystallography and NMR spectroscopy to obtain a comprehensive view on the world of protein structures.¹⁰ This tremendous increase in experimentally resolved protein structures will be accomplished by an even larger number of protein structure models computed by

homology modeling.^{11,12} The challenge, once this entire body of structural knowledge has been produced, is to extract relevant information about the properties and the functional role of individual proteins detected in particular organisms. This new strategy that seeks for the spatial structure of a protein prior to the knowledge of its actual function might provide the challenging opportunity to identify new proteins as potential drug targets.

These developments call upon methods to infer protein function directly from 3D structure. The geometry of a protein usually carries information about its biochemical function on a molecular level, e.g. as a serine protease or an oligonucleotide binding protein. However, its influence on the biological function of a cell or even more on an entire organism can only be determined in a comprehensive study where a complete series of different experimental evidences are brought together. Nevertheless, even to infer function in its biochemical sense is not straightforward, since protein

E-mail address of the corresponding author: klebe@mail.uni-marburg.de

function is not necessarily confined to a particular fold and it is often enough not apparent at the sequence level.¹³ Protein function, in particular of enzymes, is often intimately connected with the recognition and chemical modification of endogenous ligands such as agonists, antagonists, effectors or substrates. This recognition usually occurs in well-characterized clefts or cavities of protein surfaces. It has been shown for enzyme active sites that over 70% of these sites can be easily detected as the largest cleft on the surface.¹⁴ For the latter class of proteins, elementary steps of a chemical reaction are proceeded that require a strictly defined spatial arrangement of molecular recognition determinants in the enzyme active site to accommodate and spatially arrest the substrates. Very similar requirements exist for the specific recognition of co-factors in proteins or the binding of endogenous ligands (e.g. the biogenic amines) in signal transduction cascades.

This idea that molecular recognition patterns may be conserved throughout the binding pockets of proteins of similar function stimulated us to develop a new method to detect relationships among proteins. The approach involves the automatic detection and extraction of putative binding sites from proteins. Subsequently, the actual constitution of these extracted sites has to be determined and translated into molecular descriptors that are not simply based on atomic coordinates of the binding-site exposed residues but on associated physicochemical properties. Finally, the thus attributed descriptors serve as a base for the mutual comparison of different binding sites. All aspects of the algorithmic development have been conceived with regard to an efficient handling of the huge data samples of protein structures. Thus, all steps of retrieval, reduction and analysis of raw data have been drafted in a way to operate automatically, avoiding manual interference. The screening of a particular binding site against a database of several thousand binding cavities allows retrieval of proteins of similar function together with possibly bound ligands. In turn, the thus achieved indirect retrieval of bound ligands or ligand portions accommodated in structurally related binding cavities or subcavities might reveal interesting suggestions on putative bioisosteric fragments of ligands. Such ideas are extremely valuable in structure-based *de novo* design of novel leads.

Several approaches have been described in the literature to detect structural and/or functional relationships among proteins. Such similarity can be classified on three levels. The earliest comparative algorithms are based on sequence information,¹⁵⁻¹⁷ such as FASTA or routines implemented into databases (such as OWL or SWISS-PROT).¹⁸⁻²¹ While high structural homology usually matches with pronounced sequence homology, the reverse that low sequence homology parallels with structural dissimilarity, is not necessarily given. Accordingly more recent

methods determine protein similarity in terms of the overall 3D-fold. Proceeding from sequence similarity searches to comparisons based on spatial coordinates requires more complex algorithms to encode protein structural information. Considering two proteins as rigid objects, a translation and rotation matrix has to be found for spatial superimposition. This can be computed purely in geometrical terms, however more reliable solutions are obtained if the coordinates are associated with some predefined properties. Algorithms considering all atomic positions are far too demanding for fast similarity searches. Therefore approximative representations are used, e.g. based on C α -atom coordinates. The scope of similarity search algorithms following these concepts ranges from distance matrix methods,^{22,23} complete common subgraph searches²⁴ and geometric hashing techniques²⁵ to genetic algorithms.²⁶⁻²⁸ Some enhanced techniques include precalculated properties in the assignment.²⁹⁻³⁴

In most of the above-mentioned techniques, the protein structures are purely described in terms of C α -atom coordinates and geometric similarities are computed with respect to distance and angular relationships, occasionally complemented by sequence, secondary structure or amino acid property information. With respect to the recognition of a ligand in its binding site, this reduction to C α coordinates appears rather crude and limiting. However, considering all protein atoms is computationally hardly tractable. Thus, a third level of similarity search techniques has been developed that focuses on smaller subregions. They provide a compromise between computational tractability and structural complexity. The programs TESS³⁵ and ASSAM³⁶ use geometric hashing or clique detection, respectively, to retrieve templates of pre-defined 3D amino acid patterns. Through this predefinition of a particular pattern, the approaches are to some extent biased with respect to that what we can expect to be retrieved. A combination of sequence comparison to detect highly conserved residues along with recursive distance matrix alignments of coordinate sets representing "centers of functional activities" retrieves a large number of common substructures in proteins³⁷ independent from a particularly selected input structure. While most of these methods search the entire protein structure for common motifs, recent developments target the binding sites only to identify similarities that could support and assist drug design. They require assumptions about the spatial location and mutual superposition of such binding sites. In some studies, this initial step is performed manually³⁸ or based on commonly bound ligands or co-factors.³⁹ Recently Stahl *et al.*⁴⁰ reported on the analysis of 176 preselected zinc metalloproteinases that have been clustered in terms of solvent-accessible surface patches assigned to different physicochemical properties using a self-organizing neuronal net. The zinc active sites could be discriminated from other

surface depressions found on these enzymes. Approaches operating independent of bound ligands have been described by Fischer *et al.*⁴¹ and Rosen *et al.*⁴² They use geometric hashing for similarity searching, but different parameters are used to describe the binding site. Fischer *et al.* use SURFNET spheres⁴³ to generate a negative image of the active site and use coordinates of this representation for searching and comparison. Rosen *et al.* first proposed the use of generic coordinates to describe putative ligand binding. They are defined as sparse critical points,⁴⁴ assigned to patches of the Connolly surface considering its local curvature. Although this approach is less accurate compared to similarity searches based on discrete atomic coordinates and performance is reduced once the critical points are assigned to properties, their ideas inspired us to follow a related concept using a reduced set of assigned coordinates, however combined with strategies usually applied in the field of similarity searches among small molecule ligands.^{45,46}

In the present paper, we describe the algorithmic development of a new concept to compare binding pockets in proteins. A new object-oriented database Cavbase, fully integrated with the receptor-ligand database Relibase^{47,48} has been developed. As input, the method uses a large data set of cavities. For this purpose, any suitable algorithm described in the literature can be used to detect and extract surface depressions. Subsequently, descriptors to encode the molecular recognition determinants of a binding site are assigned. A clique detection algorithm is used to compare these binding site descriptions, together with a sophisticated ranking of the obtained multiple solutions. Finally, a representative set of example problems is used to assess the scope and demonstrate the power of the present method in particular with respect to data mining.

Theory and Algorithms

Cavity extraction and descriptors for cavity properties

In the present study, the comparison of a large binding site sample is attempted. To achieve this objective, special requirements with respect to the definition of binding-site regions, the assignment of reliable descriptors and the subsequent processing of the retrieved information has to be met. A number of programs have been developed to locate depressions on protein surfaces as putative binding sites. They apply different algorithmic concepts, such as flood filling techniques,^{43,49} grid-based^{50–52} or alpha shape-based approaches.^{53–56} Usually these programs operate on raw PDB data and produce either graphical output or new flat-file information. Facing the automatically extracted binding sites with areas known to bind a ligand reveals convincing agreement and underlines the

reliability of these programs. However, to avoid significant pre- and post-processing we decided to access directly the pre-processed data stored in the object-oriented database Relibase.^{57,58} We implemented the cavity detection algorithm of Ligsite⁵¹ into Relibase. Additional information from Relibase was used to attribute appropriate cavity descriptors. The accordingly extracted information has been deposited with the new database module Cavbase, sharing similar architecture with Relibase. It has been equipped with a graphical interface for data evaluation. In the Ligsite algorithm, the protein under consideration is embedded into a regularly-spaced Cartesian grid (0.5 Å grid spacing). Any grid points, represented by 1.5 Å probe spheres, penetrating into protein atoms within their van der Waals radius are discarded as solvent-inaccessible. In order to determine which solvent-accessible grid points fall into a cavity, Ligsite scans along the three Cartesian axes and the four cubic diagonals for regions, that terminate the scan directions on either ends by protein atoms. A counter is set to the number of scan directions terminated by protein atoms. This counter is used as a measure for the burial of the solvent-accessible grid points. It spans a range from 0 (fully solvent exposed) to 7 (deeply buried).

A protein must have a least one cluster of adjacent grid points comprising more than 320 grid points (approximately 40 Å³) with a degree of burial ≥ 4 . If no cluster of this size could be detected, we reduced the degree of burial for grid points to be considered in the cluster to values of three or two subsequently. This allows us to detect more shallow cavities. If present, neighboring grid points are merged into such starting clusters. The size threshold of a thus obtained cluster to be accepted has to be greater than 40 Å³. This allows it to accommodate at least one water molecule.

All surface-contacting grid points of a cluster, apart from those oriented towards the solvent, are used to approximate the cavity surface. If one atom of an amino acid residue falls closer than 1.1 Å to a protein surface-contacting grid point, the amino acid is classified as a cavity-flanking residue. These data are used to represent the basic geometric shape of cavities in the database.

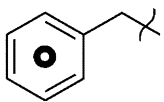
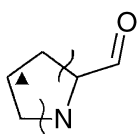
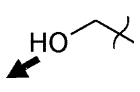
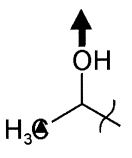
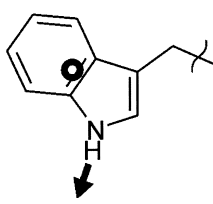
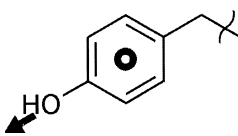
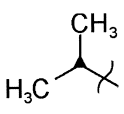
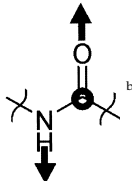
To compute spatial similarities of cavities across a large sample set an algorithm based on a restricted number of input coordinates is required. Thus, considering all coordinates of every cavity-flanking residue would be intractable. For the same reason, most approaches in the literature operate on reduced spatial information, e.g. C α coordinates instead of entire residues (see above). This strategy completely neglects the type of interactions a particular residue could possibly perform to an accommodated ligand. As a consequence, we decided to condense the physicochemical properties of the cavity-flanking residues into a restricted set of generic pseudocenters corresponding to five properties essential for molecular recognition: hydrogen-bond donor (DO), acceptor

Table 1. Encoding of the rules for the pseudocenter assignment

Side-chain	Amino acid	Pseudocenter (type)	Origin atoms
	Ala	Aliphatic	CB
	Arg	Aliphatic Donor Donor Donor	CB, CG, CD NE NH1 NH2
	Asn	Acceptor Donor	OD1 ND2
	Asp	Acceptor Acceptor	OD1 OD2
	Cys	Aliphatic	CB, SG
	Gln	Acceptor Donor	OE1 NE2
	Glu	Acceptor Acceptor	OE1 OE2
	His	PI DON_ACC DON_ACC	CG, ND1, CD2, CE1, NE2 NE1 NE2
	Ile	Aliphatic	CB, CG1, CG2, CD1
	Leu	Aliphatic	CB, CG, CD1, CD2
	Lys	Aliphatic Donor	CB, CG, CD, CE NZ
	Met	Aliphatic	CB, CG, SD, CE

(continued)

Table 1 Continued

Side-chain	Amino acid	Pseudocenter (type)	Origin atoms
	Phe	PI	CG, CD1, CD2, CE1, CE2, CZ
	Pro	Aliphatic	CB, CG, CD
	Ser	DON_ACC	OG
	Thr	Aliphatic DON_ACC	CD2 OD1
	Trp	PI Donor	CG, CD1, CD2, NE1, CE2, CE3, CZ1, CZ3, CH NE1
	Tyr	PI DON_ACC	CB, CD1, CD2, CE1, CE1, CZ OH
	Val	Aliphatic	CB, CG1, CG2
	Pep	Acceptor Donor PI	O N C

Particular atoms or functional groups define the coordinates for the different pseudocenters. Arrows indicate the origin and directionality for the mean H-bonding property exposure (v , see Figure 1) in the case of donor, acceptor and mixed donor/acceptor pseudocenters. Triangles and rings show the location of aliphatic and pi centers, respectively. The latter centers expose their properties above and below a best plane through the atoms of corresponding functional group. Aliphatic centers can be shifted towards atoms that are more exposed to the molecular surface of the cavity (see the text).

^a Thiol or disulfide bridge.

^b Peptide bond.

(AC), mixed donor/acceptor (DA, e.g. hydroxyl groups or side-chain nitrogen atoms in histidine), hydrophobic aliphatic (AL) and aromatic (PI) contact. This crucial assignment of pseudocenters to the individual amino acids obeys the rules summarized in Table 1.

The phenyl groups in Phe and Tyr are described by one PI center, respectively, representing the center-of-mass of the six ring carbon atoms. Similarly, PI centers are generated using all ring atoms in the side-chain of His and Trp, respectively. The oxygen atoms of hydroxyl groups in Ser, Thr and

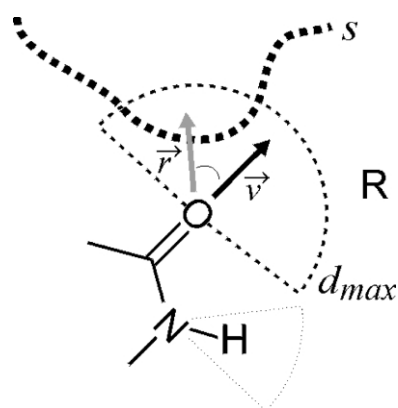


Figure 1. Two vectors, v and r , define the exposure of a particular physicochemical property. The standard vector v represents the mean direction of its property exposure, which matches in the case of an AC center, assigned to the carbonyl oxygen, the vector along the projected C=O axis. r is derived as the normalized summation vector of all vectors oriented from the oxygen to all neighboring surface grid points (S) within a predefined distance (d_{max}). The angle between v and r (Table 2) is taken as a criterion whether a pseudocenter exposes its property to a putative ligand in the cavity or whether it is removed from the list of cavity defining pseudocenters. In this example, the DO center for the backbone nitrogen would be removed.

Tyr can act as hydrogen-bond donor, and *via* their lone pairs as acceptor. The atomic coordinates of hydroxyl oxygen atoms are therefore assigned to the mixed donor/acceptor property (DA type). A similar situation holds for nitrogen atoms in His residues. The assignment of protonation states is difficult based on X-ray data in particular if the pK_a of a functional group is about 6.5 as for His. Again, the assignment of DA pseudocenters to the atomic coordinates of the two nitrogen atoms is anticipated as best compromise. Furthermore, the protonation states of the carboxy groups in Glu and Asp side-chains are sometimes difficult to define. However, on a first glance in our approximate model, these oxygen atoms are assumed to display AC centers. Peptide bonds are represented by three different types of centers, AC, DO, and PI for carbonyl oxygen, nitrogen, and carbonyl carbon, respectively. PI centers are assigned above and below a local best plane through the atoms of the peptide bond. A similar assignment would also be justified for the atoms of a terminal carboxy, carboxamide, and guanidino group in the side-chains of Asp, Glu, Asn, Gln, and Arg, however the present version neglects such assignments of PI centers. As further approximation we neglect presently hydrogen-bonding properties of sulfur atoms in Cys and Met. They are described similarly to aliphatic carbon atoms. Aliphatic centers (AL) are attributed to the side-chains of Ala, Arg, Cys, Ile, Leu, Lys, Met, Pro, and Val according to the centers-of-mass formed by their aliphatic carbon (and sulfur) atoms. According to this pro-

Table 2. Cut-off values for the angles between v and r (see Figure 1)

Pseudocenter (type)	Cut-off (°)
Donor	100
Acceptor	100
Donor/Acceptor	120
PI	60

Pseudocenters with higher values are discarded from the set of pseudocenters that define the cavity.

cedure the assigned coordinates of AL centers can vary significantly with the side-chain length and adopted conformation. This could possibly result in unreasonable spatial positions of the pseudocenters and therefore bias the contribution of the aliphatic properties in an unreasonable manner. To focus more strongly on the contributing part of surface-exposed aliphatic side-chains, only those carbon (and sulfur) atoms are considered in the calculation of the geometric mean that expose their property onto a surface area greater than 1 Å (approximately five grid points) within a distance of 3.5 Å. This reflects the scope of aliphatic interactions. The contribution of each considered atom to the geometric mean is finally weighted according to its distance from the closest surface-contacting grid points. As a result, AL centers are shifted in direction towards those aliphatic side-chain atoms that are placed next to the cavity surface. In this area they contribute most to the exposed aliphatic property. Following these rules all atoms of the cavity-flanking residues are converted into generic pseudocenters. They express the features of the 20 different amino acids in terms of five well-placed physicochemical properties.

Subsequently, the assigned pseudocenters are examined with respect to their surface exposure. This step tries to verify whether a particular interaction property could possibly form an interaction to a bound ligand. To assess their favorable exposure, the angle between the following two vectors v and r , assigned to each pseudocenter, is computed.

The first vector v describes the mean orientation along which a particular interaction could be formed. To retrieve information about given orientational preferences, data stored in the IsoStar database⁵⁹ have been consulted in detail. For example, for a DO pseudocenter, generated at the position of a nitrogen connecting two carbons, the vector v orients along the assumed NH bond. For PI centers two vectors v are generated perpendicular to the plane defined by the atoms contributing to the PI center. The pseudocenter attributed to the position of a terminal oxygen acceptor AC is represented by a vector v oriented along the projected C-O bond axis. Next, a second vector r is computed as normalized summation vector of all vectors that point from a particular pseudocenter to all neighboring surface-contacting grid points that fall into a 3 Å sphere around this center. This

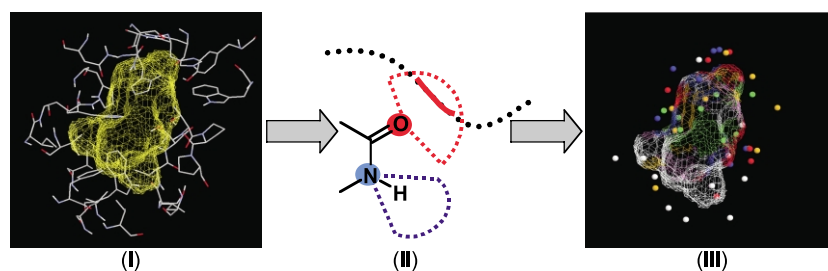


Figure 2. The shape and the properties of a binding site are determined by the amino acids (I) flanking the site. These amino acids are translated into a set of pseudocenters (III), displayed as colored spheres. Every pseudocenter exposes its property onto a certain surface patch (III). The following color scheme is used: H-bond donor (blue), H-bond acceptor

(red), ambivalent donor/acceptor (green), hydrophobic aliphatic (white) or aromatic (orange). Only those pseudocenters are considered that expose their physicochemical properties onto the surface (II).

roughly describes how the cavity surface nestles against a particular pseudo center. The angle enclosed by the two vectors serves as a criterion whether a particular pseudocenter is considered in the analysis or discarded (Figure 1). Table 2 summarizes the cut-off values for these angles. They were calibrated according to populated areas found in IsoStar.⁵⁹ AL centers are not examined with respect to these directionality criteria, mainly because they are assumed to interact isotropically in space through van der Waals forces. Finally, all surface-contacting grid points describing the cavity surface are attributed to adjacent pseudocenters. In this step, at first a distance criterion of $\leq 3 \text{ \AA}$ has to be matched and second the above-described angular selection criterion must be met. Once assigned, the various surface-contacting grid points are ascribed to one of the five physicochemical properties represented by the adjacent pseudocenter (Figure 2). According to this procedure the properties of all pseudocenters are projected onto the cavity surface. In case that surface points fall next to more than one pseudocenter within 3 \AA , assignment to the closest center is accomplished. As final result, each Ligsite-extracted surface depression is represented by a set of residue-attributed pseudocenters. The cavity surface, approximated by the set of surface-contacting grid points

is decomposed into surface patches assigned to one of the five physicochemical properties exhibited by the most adjacent pseudocenter. This abstracted description represents the input data for the cavity comparison.

Similarity searching algorithm

The detection of a common motif in two cavities, represented by the above-defined descriptors, corresponds to the problem of finding a complete common subgraph in two sets of descriptors. Solutions to this problem are discovered by clique detection algorithms.⁶⁰⁻⁶²

A 3D arrangement of pseudocenters can be regarded as a graph C for which the nodes ($c \in C$) correspond to pseudocenters and the edges correspond to distances between two pseudocenters ($d(c_i; c_j)$, with $i, j = 1, \dots, |C|$). Given a pair of graphs A and B , i.e. by nodes ($a \in A$ and $b \in B$) and edges ($d(a_i; a_j)$ and $d(b_k; b_l)$) that describe two cavities A and B , a new graph G can be defined according to the following protocol: (1) construct pairs between nodes taken from A and B , in such a way that the nodes (a_i and b_k) correspond to the same property, i.e. allowed combinations are DO_i-DO_k , AC_i-AC_k , etc. Pseudocenters assigned to the mixed property DA can also form pairs with DO

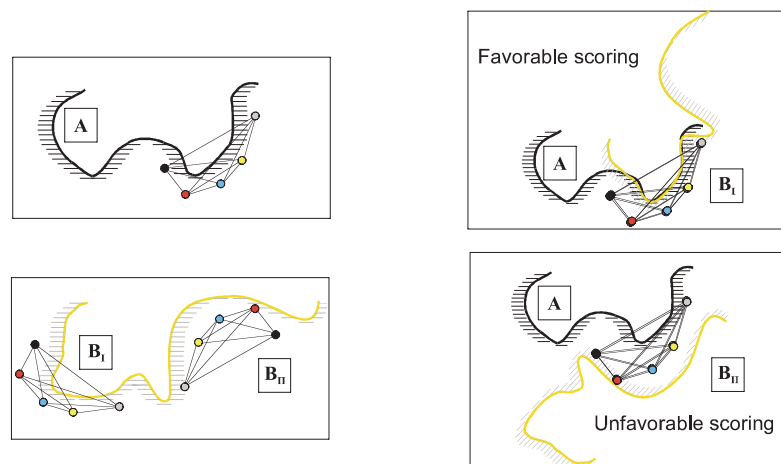


Figure 3. Schematic representation to explain the matching and scoring scheme of two binding pockets. Cavity B shares two contiguous subsets (B_I and B_{II}) of pseudocenters in common with a subset in cavity A. Either the types or distances among the pseudocenters are similar. Therefore both subsets will be detected by the clique algorithm, which is solely based on distance and property information. To determine, which substructure pattern match possesses physicochemical relevance, the corresponding pseudocenters are superimposed and a score is calculated which takes the mutual overlap of the binding site surface patches into account.

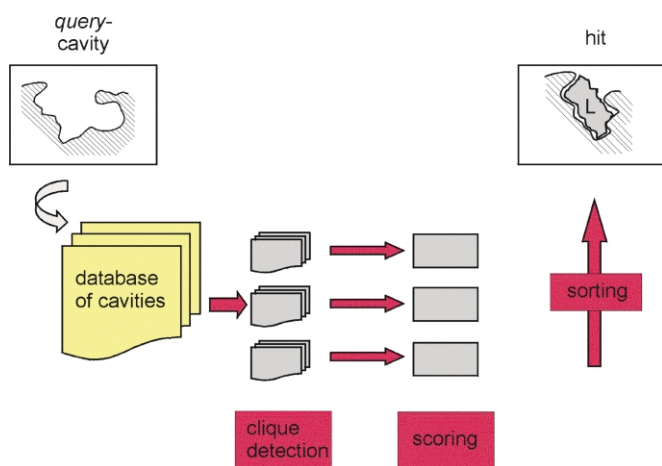


Figure 4. A cavity of interest (query cavity) is compared against a database of probe cavities. For each mutual comparison 100 clique solutions are generated and scored according to the overlapping surface patches. The best-scored clique solution of each individual comparison is kept and finally all best-scored solutions are sorted to detect the best score among all generated comparisons.

and AC centers, thus additional combinations such as DO_i-DA_k , AC_i-DA_k , *vice versa* and DA_i-DA_k are allowed. The new nodes $g_{i,k}$ in G correspond to all allowed pair combinations $(a_i; b_k)$. (2) Edges in G are defined as pairs of nodes $(g_{i,k}; g_{j,l})$ for which the actual distances among the generic pseudocenters match within a predefined tolerance, i.e. $g_{i,k}$ and $g_{j,l}$ are connected if $d(a_i; a_j) \approx d(b_k; b_l)$. The distance tolerance has been set to 2 Å to cope for spatial uncertainties in the pseudocenter positions. This value has been rationalized by comparing multiple PDB entries of the same protein however bound to different ligands. It has to be remembered that such deviations originate either from the limited accuracy of protein crystal structure determinations and, even more pronounced, from conformational differences in the cavity-flanking residues among related proteins. In addition, only intercenter distances up to 12 Å have been regarded. Rationale behind this cut-off is the limitation of our similarity search to distances in the short and medium range, in particular since the described uncertainties are likely to increase at longer distances.

The Bron-Kerbosh algorithm⁶⁰ has been applied to find the maximal common subgraph in G . Reflected back onto the pseudocenters, such a common subgraph represents a similar spatial arrangement of properties in two cavities, thus defining a similar motif. Clique detection algorithms are computationally demanding, since they scale with N^2 for every additional node $g_{i,k}$. However, the above-described assignments produce a limited set of descriptors being a satisfactory compromise between required accuracy (number of centers) and computational tractability (pair-wise comparison of about 70 centers takes approximately three seconds on a state-of-the-art Linux processor). Solely considering the distance matrix among pseudocenters can still produce chemically unreasonable solutions. E.g. the result from a clique detection is geometrically still reasonable if equivalent centers from a concave area in one cavity match upon a convex one in a second (Figure 3).

In such cases, the actual superposition reveals a chemically unreasonable match. Obviously, the direction of property exposure matters in the comparison. Attempts to consider directionality in the definition of nodes $g_{i,k}$ by means of the vectors v and/or r , did not improve the detection of correct solutions. Thus, we decided to compute an independent scoring to rank the generated clique solutions. It considers the mutual matching of assigned surface patches of the two cavities aligned, according to the shared pseudocenters detected by the clique algorithm. The following protocol is accomplished (Figure 4): For each pairwise comparison of cavities, the 100 largest common subgraphs are evaluated. This results in 100 common spatial arrangements of pair-wise matching pseudocenters (individual clique solutions).[†] Each clique solution generates a matrix that transforms the matching pseudocenters together with the associated surface patches onto those of the reference as best spatial superposition. Subsequently, each generated superposition is analyzed and the overlap in surface points ($p \in P$), assigned to the same physicochemical property, is determined. These surface points originated from the embedded grid of 0.5 Å spacing, thus the mutual overlap S_{AB} of points from the two matching surfaces is calculated as:

$$S_{AB} = \sum_v \sigma_v \quad (\text{for all } \sigma_v \geq 0.7)$$

$$\sigma = \frac{\rho_{ai} + \rho_{bk}}{|P_{ai}| + |P_{bk}|}$$

with

$$\rho_{ai} = |\{p_{ai} | d(p_{ai}; p_{bk}) \leq 1.0\}|$$

[†] A value of 100 is the best empirically determined compromise between computational effort and achieved coverage of solutions.

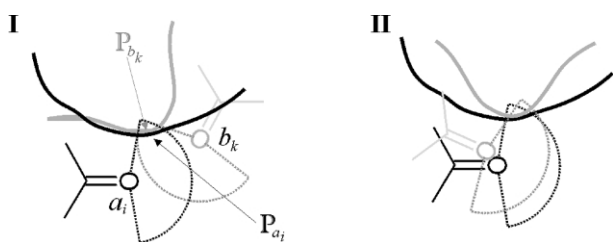


Figure 5. Schematic drawing of two possibilities to superimpose minor surface patches P generated by the Acceptor centers a_i and b_k of two carbonyl groups. Both superpositions result in very similar surface matches and will therefore contribute equally to S_{AB} . Thus R_1 cannot discriminate between both solutions, R_2 , however, will favor solution II due to an implicit consideration of the directionality *via* the *RMSD* value.

and

$$\rho_{bk} = |\{p_{bk} | d(p_{bk}; p_{ai}) \leq 1.0\}|$$

The degree of mutual overlap in surface patches is expressed by the number of surface points that fall next to each other below a distance threshold of 1.0 Å. To avoid consideration of strongly fragmented surface patches, the mutual overlap of patches is only counted if at least 70% of the matched surface patches†, corresponding to a pseudocenter pair, fall next to each other below 1 Å. Accordingly, if n common pseudocenter pairs have been detected in the subgraph analysis, S_{AB} exceeds maximally to a value of n and minimally to a value of $S_{AB} = 0.7n$. Those of the pseudocenter pairs that pass the overlap criterion of adjacent surface patches ($\geq 70\%$) define a subset of equivalent centers representing the physicochemical properties shared by both cavities. In a subsequent refinement step, a new transformation matrix is computed, however, only considering those pseudocenters that passed the above-defined overlap criterion. Finally, a new ranking is calculated by determining the matched surface points approaching each other below 1 Å after the second transformation has been performed.

This procedure is followed for the above-mentioned 100 best clique solutions in the pairwise cavity comparison. The procedure reveals 100 improved solutions with pairs of equivalent pseudocenters. Out of these, the solution with the highest S_{AB} value is stored together with the corresponding set of n equivalent pseudocenter pairs. On the mean, the refinement and scoring procedure requires additional 100 seconds for two medium sized cavities (ca. 800 Å³) on a state-of-the-art Linux processor.

For each cavity in a test set such a pairwise comparison with a query cavity is performed. The various solutions, obtained for the entire sample of

test cavities, are ranked according to S_{AB} and n . The query cavity to be compared with the test sample could comprise all pseudocenters representing the entire binding site or could be reduced to a pseudocenter subset, e.g. to a specifically edited sub-pocket. Two figures-of-merit (R_1 and R_2) are considered to rank the entire set of cavities with respect to their similarity with the query cavity:

$$R_1 = S_{AB}$$

and

$$R_2 = \frac{S_{AB} - 0.7n}{RMSD}$$

where *RMSD* corresponds to the root mean square deviation of the matched pseudocenter pairs used for superpositioning. Using R_1 , the list of pairwise comparisons is simply sorted in terms of the size of their achieved surface overlap, accordingly cavities that share multiple surface patches in common with the query cavity will occur at the top of the list. Visual inspection of the top-ranked cavity matches disclose some deficiencies while focusing entirely on R_1 . In particular fragmented and disconnected motifs of rather small surface patches produce a mutual overlap that is not very conclusive with respect to shared property distributions. Figure 5 illustrates such a situation where two possible superpositions with similar σ and accordingly S_{AB} result in equivalent R_1 values, even so the match of pseudocenters is quite unsatisfactory (Figure 5, right). Obviously, the pure summation over common surface patches has to be weighted by the total number of contributing pseudocenters and their spatial matching accuracy. Thus, we rank their mutual match with respect to the spatial deviation (*RMSD*) achieved in the cavity superpositioning step. The term $S_{AB} - 0.7n$ describes the relative size of the overlapping surface patches, since S_{AB} can adopt maximally a value of n . Any considerations to include vectors v and/or r as directionality terms in the scoring and ranking did not improve the figures-of-merit. In practice, we sort our comparisons according to R_1 and during visual inspection we consult on purpose R_2 , S_{AB} and n . Therefore Cavbase has been equipped similarly to Relibase, with a visualization tool based on RASMOL.⁶³ It allows one to display and browse through the top-ranked hits of the similarity analysis in short time.

Prefiltering of data sample

The Ligsite approach used in this study automatically extracts depression on the protein surface as putative binding site. However, not necessarily all of them are relevant on a first glance. This could possibly intricate a critical assessment of the performance of the new method. Accordingly, for our initial validation we considered only cavities that accommodate at least one small molecule

† This value is an empirical estimate that does not penalize conformational deviations too strongly, but considers similar property exposure onto the surface.

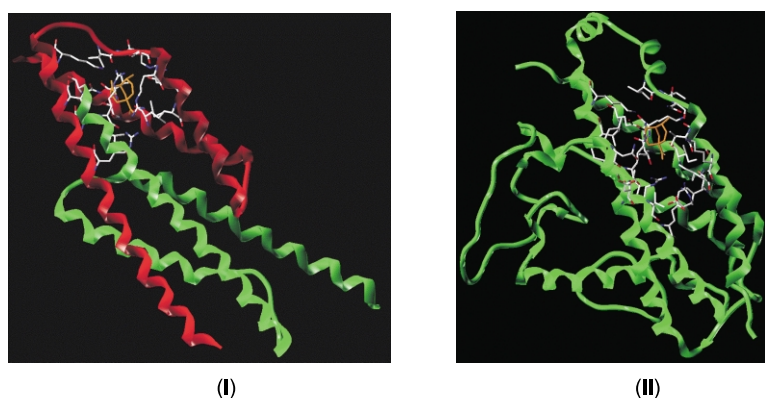


Figure 6. Folding pattern and amino acids comprising the binding sites of chorismate mutases 1ecm (*E. coli* (I)) and 4csm (*S. cerevisiae* (II)) are shown together with the bound ligand. In the first case, the binding site is composed of two different polypeptide chains, whereas in the latter case the amino acids originate from one contiguously connected chain.

with six up to 50 non-hydrogen atoms, thus covering the range typically found for small drug molecules.

A further complication arises from redundancies present in the stored PDB data. Multiple cavity entries will occur in Cavbase originating from distinct PDB entries, however extracted from the same underlying protein. In a comparative analysis these cavity entries will produce a trivial high similarity score. A possible filter to eliminate such redundancies in the considered cavity data set could be a pre-selection of PDB entries that are confirmed as highly diverse with respect to their overall 3D structure, e.g. similar to the selection performed by Fischer *et al.*⁶⁴ However, we refrained from pre-selecting PDB entries on the base of folding patterns, since we wanted to avoid any selection based on overall protein information, e.g. folding patterns. Furthermore, the consideration of high sequence similarity as pre-filter is also not fully reliable since the cavity flanking residues not necessarily originate from one contiguously connected peptide chain (Figure 6). Often enough several domains contribute to a cavity and a decision would have to be taken which peptide chain(s) to consider for sequence comparison. However, comparisons on the sequence level would not conflict with the precondition that our approach should neglect any structural information apart from the spatial composition of physicochemical properties exposed to a binding site. Therefore, the overall cavity data set is divided into clusters with expected trivial internal similarities pursuing the following protocol:

(1) Each cavity is analyzed with respect to the peptide chains that contribute cavity flanking residues.

(2) For each thus detected peptide chain, its corresponding representative chain is retrieved from the PDB-select database.^{65,66} In this database, to each peptide chain found in the PDB, a representative chain has been assigned according to an all-against-all sequence comparison. Thus, for a particular cavity under investigation, a set of one or more representative chains is attributed resulting in a set of parent sequences.

In case that several chains contribute cavity flanking residues and all share the same representative chain in common in the PDB-select database, multiple assignments with the parent set can occur.

(3) Cavities assigned to the same parent set according to (1) and (2) are then clustered together. Thus, the procedure groups cavities together that are composed of peptide chains with high sequence similarity. Cavities falling into the same cluster will likely possess a trivial similarity score. To avoid such trivial comparisons in our study, we only use one representative cavity from each such formed cluster as query cavity. Subsequently, it is compared to all other cavities in the remaining clusters.

Results

Results from the pre-filtering process

The automatic extraction of binding pockets has been applied to the June 2000 version of the PDB containing 11,983 entries. Of these, 8627 with a resolution of 3 Å or better have been considered after discarding all model-built structures, NMR and superseded entries or data containing only C α coordinates.

A sample set of 31,441 surface depressions could be detected using our implementation of the Ligsite algorithm. These were stored in the new object-oriented Relibase module Cavbase.

For the validation of our method, we considered only PDB entries that contain at least one ligand with 6 up to 50 non-hydrogen atoms. The definition of a "ligand" thereby obeys the rules set in Relibase. Accordingly, 4332 PDB entries remained, corresponding to 18,402 cavities. This data sample was further analyzed to consider only cavities that actually accommodate a ligand. With the precondition that at least one ligand atom must be buried, only 5448 cavities remained originating from 3626 independent PDB entries. In 696 cases, ligands (mainly sugars) coincide with flat surface regions

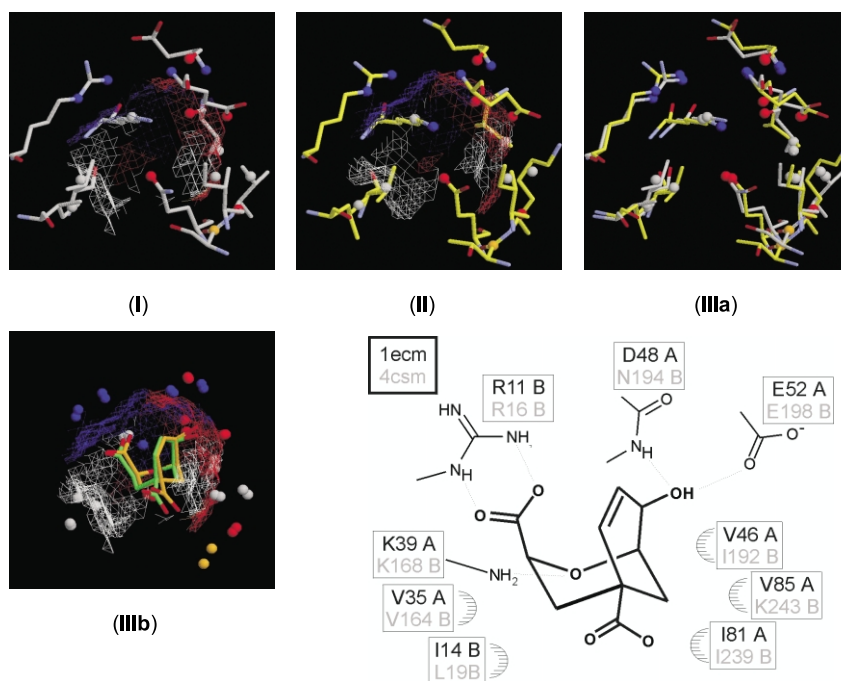


Figure 7. The binding site sketch illustrates that the algorithm can handle equivalences originating from different chains. While in the case of 4csm (II) the cavity is entirely built-up from a single chain (B), the active site interactions in 1ecm (I) reside from two different protein subunits (A and B). (III) shows the superposition of the involved amino acids (IIIa) and the surface patches (IIIb). The example illustrates the importance to use generic descriptors, since equivalent H-bonding and hydrophobic properties are not necessarily experienced by one particular type of amino acid.

that are not detected as cavities or the ligand could be classified as solvent molecule (e.g. benzene) with more than five atoms. A statistical evaluation of the volumes found in the data sample of the 18,402 extracted cavities shows that most cavities fall into a range between 300 and 800 Å³. Regarding only the subset of cavities actually occupied by a ligand reveals a distribution between 300 and 1500 Å³ with a significant shift of the mean towards larger volumes (mean about 800 Å³). Obviously, many of the small pockets (<500 Å³) remain unoccupied.

One might expect that cavities retrieved from distinct PDB entries, however originating from the

same underlying protein, possess very similar shape and thus property distributions. Due to conformational flexibility of proteins, cavities of quite deviating size can be extracted. To some extent this is also a result of the parameter settings applied in Ligsite. Small conformational changes of the protein found in different PDB entries can trigger the detection of subpockets or even extended channels in one entry that is inaccessible in the other. This results in significant size and shape differences of cavities although originating from the same parent protein. The clique detection algorithm used in our approach for the comparative analysis is capable of coping with such size differences since it seeks for

Table 3. Equivalent pseudocenter pairs and the involved amino acids of the chorismate mutases structures 1ecm and 4csm used in the cavity matching algorithm.

Type of equivalent pseudocenter pairs	1ecm (<i>E. coli</i>)		Corresponding amino acids ^a		4csm (<i>S. cerevisiae</i>)	
Donor	K39	A	s	K168	B	s
Acceptor	V46	A	p	I192	B	p
Donor	D48	A	p	N194	B	p
Acceptor	D48	A	p	N194	B	p
Donor	E52	A	p	E198	B	p
Acceptor	E52	A	s	E198	B	s
Acceptor	I81	A	p	I239	B	p
PI	S84	A	p	T242	B	p
Acceptor	Q88	A	p	D246	B	s
Aliphatic	V35	A	s	V164	B	s
Aliphatic	K39	A	s	K168	B	s
Aliphatic	V46	A	s	I192	B	s
Aliphatic	I81	A	s	I239	B	s
Aliphatic	V85	A	s	K243	B	s
Donor	R11	B	s	R16	B	s
Donor	R11	B	s	R16	B	s
Aliphatic	I14	B	s	L19	B	s

^a One-letter residue name, residue number, chain-ID and origin (s: center originates from side-chain atom(s); p: center originates from backbone atom).

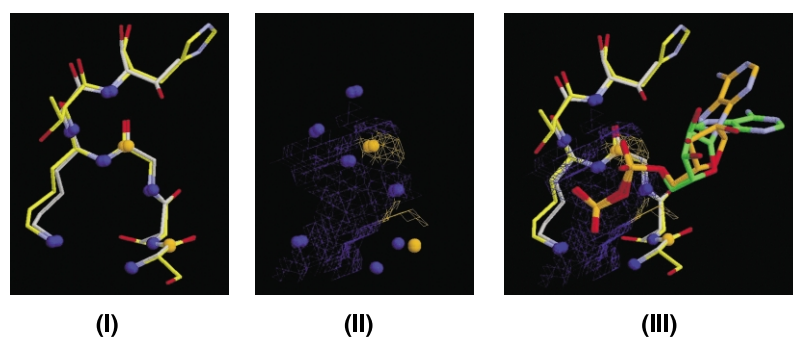


Figure 8. Equivalent phosphate binding areas in the binding pockets of uridylyl kinase (1ukz) and the structure of a kinesine-type domain (3kar). The superposition based on the matching pseudocenters shows extensive conservation of the involved amino acids (I) and the common surface patches (II). The phosphate groups of the bound ADP-ligands superimpose well (III).

common subgraph matches. Finally, the dataset of 5448 cavities has been further clustered into 1010 parent sets from which the query cavities were selected and subsequently compared against the remaining set of probe cavities.

Validation of the cavity matching algorithm

In the following we will use a representative set of bench mark examples, partly taken from literature, to assess and demonstrate the scope, applicability and success rate of our new approach. As a first test, we retrieve the binding pocket from the same protein present in different species sharing low sequence homology. We then move to the detection of binding sites accommodating the same ligand. Finally, binding sites are compared that catalyze similar chemical reactions.

Similarity between binding pockets of two chorismate mutases originating from different species

As a first comparative example we selected two binding pockets extracted from chorismate mutases originating from two species: *Saccharomyces cerevisiae* and *Escherichia coli*.^{67,68} This example has previously been studied by Rosen *et al.*⁴² using sparse critical points⁴⁴ derived from the Connolly surface of previously extracted binding pockets. In a one-to-one comparison they could successfully detect a common surface-point pattern in the two binding pockets. Although the two proteins show a sequence identity of less than 20% they adopt a similar fold and bind the same ligand. The bicyclic transition-state-analog inhibitor is recognized in both cases *via* side-chain interactions. In order to examine whether our approach is capable to retrieve and match the two chorismate cavities, we defined the cavity from *S. cerevisiae* as query and screened our complete sample set including the chorismate example from *E. coli*. Both scoring criteria R_1 and R_2 place the *E. coli* cavity on the best rank. The actually obtained match is shown in Figure 7. Table 3 lists the corresponding pseudocenters with the associated amino acids used in the superpositioning procedure. Although the actual coordinates of the ligands were not used in the approach, the obtained cavity surface match generates a trans-

formation that displays the bound inhibitors in a virtually perfect superposition. Even so the enzymes show similar fold, ligand recognition does not necessarily require the same amino acid composition of the binding site. This result is particularly remarkable since the actual interactions to the ligands are nearly exclusively performed by side-chain contacts. Furthermore, in the *S. cerevisiae* enzyme the binding pocket is composed by residues emerging from one peptide chain whereas in the *E. coli* protein two chains are contributing (see Figure 6). This fact would clearly limit the applicability of sequence alignment methods to detect cavity similarity.

Similarity between portions of co-factor binding pockets in non-homologous proteins

The successful retrieval of two pockets recognizing the same rigid ligand prompted us to extend our approach to a larger set of similar and more flexible ligands. Cofactors are frequently found as common ligands in proteins, accordingly their binding has been matter of comparative studies. Already in 1984, Hol & Wierenga⁶⁹ detected common binding-site features next to phosphate groups of bound ligands. Extended α -helical structural motifs generate a partially charged, highly polar binding region favorably occupied by negatively charged phosphate groups. Kinoshita *et al.*⁷⁰ studied the local environment of phosphate groups in nucleotide-binding proteins *via* the comparison of all protein atom coordinates in a sphere of 7 Å. Obviously structural similarity has been detected. Usually several NH groups either of the backbone or side-chains point towards the phosphate binding site. In our approach we selected the local phosphate recognition site of a kinesine-type domain (3kar)⁷¹ accommodating ADP and queried the observed pattern against other proteins. The cavity from a uridylyl kinase (1ukz)⁷² shows high local similarity giving rise to the mutual alignment presented in Figure 8. Actually this latter enzyme exhibits no sequence and fold similarity with the protein from where the query cavity had been extracted. Nevertheless, it similarly recognizes the phosphate position of an ADP.

As another example for local pattern matching, Kobiyashi *et al.*³⁸ and Moodie *et al.*³⁹ investigated on the basis of well selected data samples the local

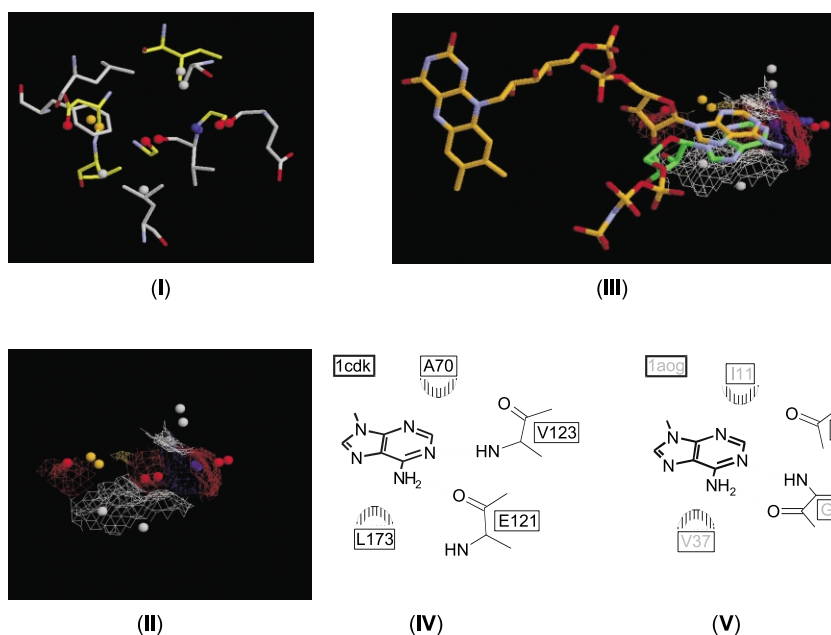


Figure 9. Pseudocenter and surface patch patterns shared in common between two adenine-binding regions. The underlying protein structures show no significant sequence and fold homology; (I) depicts the superposition of the matching pseudocenters of 1cdk and 1aog (yellow or white carbon atoms, respectively) and (II) the superposition of the corresponding surface patches. In (III), the bound ligands are also displayed being 5'-adenyl-imido-triphosphate (1cdk, carbons in green) and flavine-adenine-dinucleotide (1aog). In (IV) and (V) the corresponding interactions with the respective adenine fragments are illustrated.

binding environment of adenine portions in their protein receptors. In the first study, a manual binding-site superposition has been performed whereas Moodie *et al.* detected a conserved pattern of physicochemical properties apart from the actual amino acid composition. In our analysis we refrained from a sophisticated pre-selection of a data sample. Instead we picked by chance the pocket from a cAMP-dependent protein kinase (1cdk)⁷³ accommodating 5'-adenyl-imido-triphosphate as ligand. The query pocket has been edited to display the local environment adjacent to the adenine portion. To avoid any trivial similarity matches all entries were removed possessing high sequence identity with 1cdk according to the assignment in the PDB-select database. The remaining 5431 entries have been used for the mutual comparison with the query pocket.

Based on the R_1 scoring, cavities have been detected on the first ranks that adopt, according to the FSSP score, the same fold as the reference,

however with no significant sequence homology. Already on rank 7, the cavity extracted from trypanothione reductase (1aog)⁷⁴ possessing no sequence and fold homology with 1cdk, is found (Figure 9). The matching pseudocenters are listed in Table 4. This example demonstrates that an extensive correspondence in surface portions or pseudocenters does not necessarily result from a close spatial alignment of the contributing amino acid residues, but more important from a firm resemblance of spatial physicochemical properties in space. Considering the actually bound ligands, a convincing spatial match of the adenine moiety in 5'-adenyl-imido-triphosphate is found with that of flavine-adenine-dinucleotide in 1aog.

Table 4. Equivalent pseudocenter pairs and the involved amino acids found in the adenine binding regions of the structures 1cdk and 1aog used in the cavity matching algorithm

Type of equivalent pseudocenter pairs	Corresponding amino acids ^a					
	1cdk		1aog			
Acceptor	L49	B	p	D36	A	s
Acceptor	E121	B	p	G128	A	p
Donor	V123	B	p	G128	A	p
Acceptor	V123	B	p	G126	A	p
PI	F327	B	s	D36	A	P
Aliphatic	A70	B	s	I11	A	s
Aliphatic	L173	B	s	V37	A	s

^a One-letter residue name, residue number, chain-ID and origin (s: center originates from side-chain atom(s); p: center originates from backbone atom).

Matching entire co-factor binding pockets

In the previous examples only binding sub-cavities recognizing recurrently similar ligand portions have been investigated. To examine the scope of our approach, we selected NADPH as a much larger ligand frequently observed in protein structures. As query cavity we extracted the pocket in a carbonyl reductase (1cyd).⁷⁵ To avoid trivial matches any proteins with high sequence homology to 1cyd have been discarded from our sample set yielding 5377 cavities. As a result from this search on the first ranks based on R_1 or R_2 only cavities from other oxidoreductases have been detected (see Supplementary Material), all accommodating either a complete NADP(H) or NAD(H) ligand. Cavities placed at minor ranks host co-factors with decreasing similarity however still containing parts of the NADPH skeleton. Approximately from rank 50 onwards only parts of the NADPH co-factor binding pocket are recognized, occasionally because the co-factor adopts a deviating conformation from that in the query

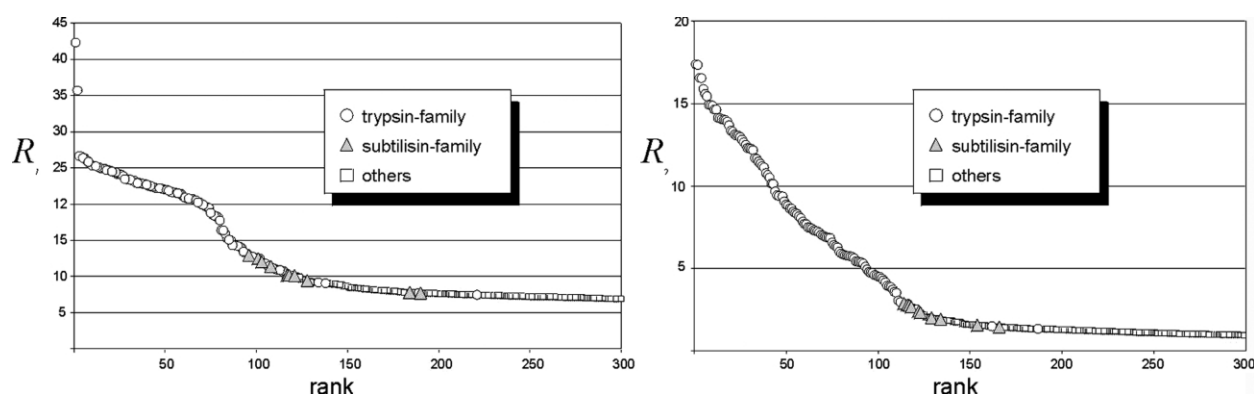


Figure 10. The first 300 best-ranked solutions from a comparison of the binding site in the trypsin structure 1tpo against a set of 5284 probe cavities are shown. Sorting has been performed according to R_1 and R_2 . The first example for a cavity from the subtilisin-family is found on rank 95 (R_1) or on rank 113 (R_2).

cavity. Furthermore, at these ranks cavities are detected hosting other ligands usually originating from proteins with different function. At rank 98, the cavity of a methyl transferase (2adm),⁷⁶ binding *S*-adenosylmethionine is found and at rank 116 the pocket of a phenol hydrolase (1foh)⁷⁷ is observed that accommodates FAD as co-factor. Here, the adenine-recognition site is shared in common with the original NADPH query pocket.

The present example provides another remarkable insight. On rank 41 the large cavity of a NADPH-dependent steroid dehydrogenase (1fds)⁷⁸ is detected, although its crystal structure has been determined in the absence of the bound co-factor. Only because a steroid is present in this large pocket, the entry remained in our data sample. Nevertheless, the co-factor cavity observed in the carbonyl reductase matches well with the large unoccupied part of the pocket in the steroid dehydrogenase and falls next to the binding site of the steroid. Thus our approach can also be used to detect and match unoccupied binding sites in a comparative analysis.

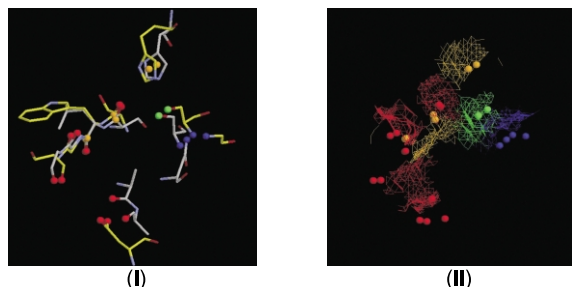


Figure 11. Superposition of the cavity from trypsin (1tpo, white carbon atoms) and from proteinase K (2prk, yellow carbon atoms) based on the pseudocenter patterns matched as similar in both cases. Obviously, from the catalytic triad the histidine, and the serine are considered together with the oxy-anion hole and the non-specific peptide recognition site. The catalytic aspartate is not surface-exposed. The superimposed amino acids (I) and surface patches (II) are shown.

Recognition of binding cavities in proteins of similar biochemical function

The previous case study convinced us that our approach should be suited to retrieve proteins of common biochemical function from an extended sample set. For our analysis we selected serine proteases. This example has previously been studied. Fischer *et al.*⁴¹ presented a geometric hashing algorithm to detect similarities among serine proteases. However, they used molecular descriptors taken from the entire protein structure for their analysis. These reflect more strongly features based on the protein fold rather than our method that reflects physicochemical properties experienced by the binding-site residues only.

In a first run, the binding pocket of the trypsin structure 1tpo⁷⁹ has been selected as query cavity. The remaining set has been reduced to 5248 entries by discarding those cases that were expected to produce trivial similarity solutions (criteria applied

Table 5. Equivalent pseudocenter pairs and the involved amino acids found in catalytic centers of two non-equivalent serine proteases from the trypsin (1tpo) and subtilisin-like family (2prk)

Type of equivalent pseudocenter pairs	Corresponding amino acids ^a			
	1tpo		2prk	
Aromatic	H57	s	H69	s
Acceptor	D189	s	A159	p
Acceptor	D189	s	A158	p
Donor	G193	p	N161	p
Donor	S195	p	S224	p
Donor/Acceptor	S195	s	S224	s
PI	S214	p	S132	p
Acceptor	S214	p	S132	p
PI	W215	p	L133	p
Acceptor	W215	p	L133	p
Acceptor	G216	p	G134	p
Acceptor	S217	p	G135	p

^a One-letter residue name, residue number and origin (s: center originates from side-chain atom(s); p: center originates from backbone atom).

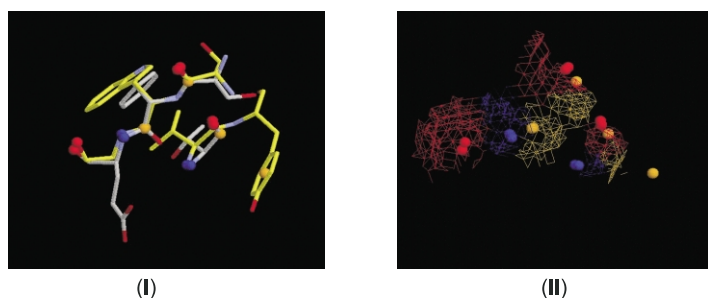


Figure 12. (I) The superposition of the amino acids of trypsin (1tpo, carbon atoms colored in yellow) and a ketosteroid isomerase (1qjg, carbon atoms in white). The ketosteroid isomerase was found on rank 120. In (II) the superposition of surface patches is shown. The matching pattern comprises several pseudocenters involved in the peptide backbone exposed to the binding pocket.

as above). The comparison was ranked according to R_1 and R_2 (Figure 10) (see Supplementary Material). On the top ranks only other members of the trypsin family were detected such as thrombin, chymotrypsin or tryptase. These are followed by other examples adopting similar fold, however exhibiting decreasing sequence similarity (e.g. kallikrein A, factor D, α -lytic protease or proteinase A). At rank 113, thus among the first 3% of the data sample considered, a binding pocket from the structurally unrelated subtilisin family (1sue)⁸⁰ has been registered. Trypsin and subtilisin are both representative parent structures of the two major serine protease classes. They share the same biochemical and mechanistic function, however without sequence and fold homology. On the following ranks other examples of the subtilisin family were found. Comparing the individual matches of pseudocenters or common surface patches among members of the two families shows that the physicochemical properties of the catalytic serine and histidine are matched (the aspartate is not surface-exposed) along with the oxyanion hole and the binding features experienced by the non-specific peptide recognition site (Figure 11 and Table 5).

Apparently a scoring based on R_2 is better discriminating, however based on R_1 members of the subtilisin family are recognized on even more prominent ranks. A critical assessment of the results should not ignore that our approach also produces solutions that appear on a first glance of no relevance for the detection of common functional features. On rank 120, a common patch is shown between the query trypsin cavity and that extracted from a ketosteroid isomerase (1qjg).⁸¹

The common pattern expands in this case over several pseudocenters assigned to atoms involved in the peptide backbone. According to the given molecular dimensions in such structural elements a common and obviously rather repetitive pattern is detected by the clique algorithm once such a unit is exposed to the cavity surface (Figure 12).

To assess the reliability of our approach we inverted the “serine-protease” query, now selecting the subtilisin binding pocket of 1sua⁸² as query cavity. In this query other subtilisin cavities were not excluded. As expected this run retrieves at first the other entries from the subtilisin family followed by proteinase K and members of trypsin-type family on the subsequent ranks. The listing of entries clearly shows that the data base is less populated of structural variants by the subtilisin family (see Supplementary Material).

Idea generator for *de novo* design

De novo design seeks for novel ligand skeletons to occupy a given binding pocket. The detection of common surface patches among binding pockets might provide some new ideas about possible lead structures *via* the analysis of the actual cavity occupants. However, it should be noted that besides the initial selection of the cavity data set, no ligand information is used in the approach. Interestingly enough we detected a surface patch of an adenine-binding pocket to be similar with an unoccupied binding-site region in HIV protease. In a study by Martin *et al.*⁸³ the binding of a series of macrocyclic peptidomimetic inhibitors to HIV protease is described. Depending on the substitution pattern these ligands orient different

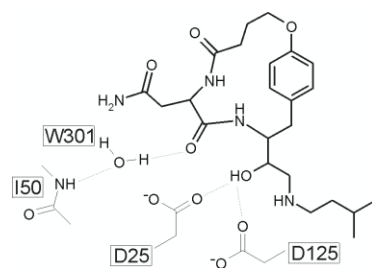
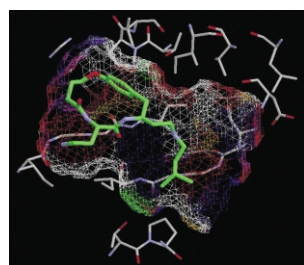


Figure 13. Binding site of the HIV-protease (1b6o) with a bound macrocyclic peptidomimetic inhibitor. For reasons of clarity some amino acids are not shown. An unoccupied area exhibiting similar physicochemical properties compared to an adenine-binding site present in 1cdk is indicated. On the right, a schematic illustration of the most important interactions formed to the inhibitor are plotted.

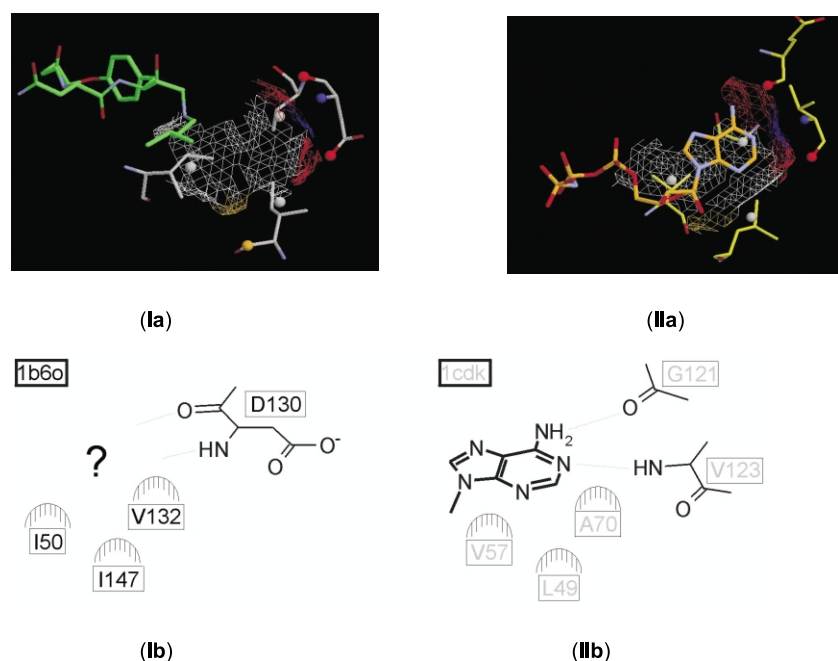


Figure 14. Area of physicochemical equivalence found in the unoccupied portion of the binding site in HIV protease 1b60 (**Ia**) and the adenine-binding cavity in 1cdk (**IIa**). The interactions matched as equivalent are shown in (**Ib**) and (**IIb**).

molecular portions into the S2' subpocket, e.g. a *p*-aminosulfonamide group. However, in the pdb entry 1b60 the bound ligand lacks a phenyl-sulfonyl group at this position, thus leaving the addressed binding-site niche unoccupied. The corresponding surface patch of this niche occurs similarly in the catalytic subunit of protein kinase A (Figure 13 and Figure 14). There the patch accommodates the adenine portion of adenylylaminophosphate. The discovery of such examples could be of potential interest to *de novo* design of protein ligands. Molecular building blocks detected by this approach are actually known to be recognized at a site with a particular protein surface pattern. Possibly they can be joint with a ligand that occupies the remaining part of the binding site. The present HIV example stimulated us to perform a more detailed search using the described binding-site niche as a query cavity. The search against a dataset of 7192 probe cavities revealed several examples besides ligands containing an adenine moiety where this patch is occupied

by a hydrophobic aromatic group being part of a larger ligand, e.g. a *p*-hydroxybenzamidyl group in the cAMP-dependent protein kinase inhibitor balanol (1bx6)⁸⁴ (Figure 15). This finding matches well with the fact that the structural studies of Tyndall *et al.* actually show in one of the HIV complexes a phenyl group filling up this niche (1d4 l).⁸⁵ These last examples demonstrate the potential use of our approach in ligand *de novo* design. Likely, a large enough database of cavities together with their bound ligands can be used to generate interesting suggestions how to modify and improve known protein ligands.

Conclusions and Outlook

As a basic concept, our new approach assumes that similar function among proteins requires similar binding pockets. These pockets have to expose spatially conserved physicochemical properties in order to recognize and subsequently respond to

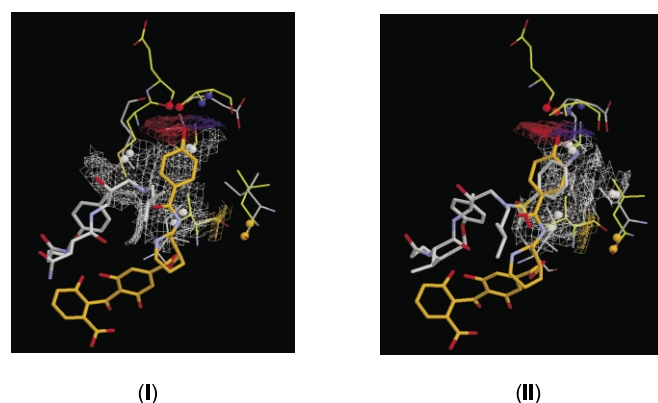


Figure 15. Superposition of the binding sites of HIV-protease 1b60 (carbon atoms in white) and protein kinase 1bx6 (carbon atoms in yellow) is given in (I). The *p*-hydroxy-benzamidyl group of the kinase ligand binds right into the hydrophobic niche where the phenyl-sulfonyl group of the 1b60 ligand is missing (see the text for further explanation). For comparative purposes the superposition of HIV-protease 1d4 l with 1bx6 is displayed in (II), showing the close overlap of the aromatic moiety in the hydrophobic niche.

the binding of the same or a related substrate or endogenous ligand. Following this idea we depart from the actual sequence or fold information and move to a more general description of features that determine conserved interaction patterns. We developed a unique coding scheme to condense the properties of cavity-flanking residues into simple descriptors. Common recognition patterns in terms of conserved subsets of these associated descriptors are detected using a clique detection algorithm. However, to retrieve relevant information a reliable scoring scheme is essential that measures surface patches shared among the matched sub-pockets in common.

Using a representative set of benchmark examples we could demonstrate the scope of our approach. It extracts and matches from a sample set of several thousand cavities extracted from non-homologous proteins those examples that recognize the same ligand. Equally well entire co-factor binding sites can be retrieved. With decreasing similarity to the skeleton of the reference co-factor, the method also matches only sub-pockets shared in common with the query cavity accommodating the reference co-factor. Functional relationships among proteins resulting in the catalysis of a similar enzymatic reaction can be retrieved with our approach. However, in our opinion the most promising aspect is its potential to suggest alternative molecular building blocks in *de novo* design. The search for putative molecular portions well-suited to accommodate a particular sub-pocket of the binding site under consideration can be inspired by the retrieval of ligands actually occupying a very similar sub-pocket in other already structurally characterized proteins. Such a source of information for ligand design may develop a routine tool in supporting structure-based drug design.

A further aspect could be of potential relevance in understanding drug action. Frequently, side effects of drugs are created due to undesired binding to the pocket of another protein. A search based on the described approach provides the possibility to detect structurally related binding cavities where such unexpected binding could occur. Important enough this prediction does not rely on the ligand properties but purely on the shape of the binding pockets. This leaves room to modify a ligand structurally to achieve better selectivity.

The present approach requires several improvements. First-of-all the classification of exposed amino acids has to cover all relevant interaction patterns. At present our approach neglects some supposedly important contact geometries (e.g. π -stacking to carboxy or guanidine groups). Secondly, the scoring of the different solutions suggested by the clique detection algorithm is essential for the retrieval of relevant information. At present it is based on surface patches shared in common by the matching cavities. Improved figures-of-merit have to consider better the con-

tiguous connection of matched surface patches. Finally, at present our approach is computationally rather intensive. It requires algorithmic accelerations. Such improvements would enable an all-against-all comparison of the entire cavity database. Such a study is likely to provide an entirely new classification and clustering of protein structures in terms of cavity similarity aspects.

Acknowledgements

The authors acknowledge stimulating discussion with Dr M. Hendlich (Lion Biosciences, Heidelberg, Germany) in particular in the beginning of this project. The help of Judith Günther (Univ. Marburg) and Dr A. Bergner (CCDC, Cambridge, UK) in implementing various aspects of Cavbase is gratefully acknowledged. We thank Professor A. Ultsch and R. Simon (University Marburg) for helpful discussions about algorithmic aspects. The present project has been supported by the German Minister of Science and Education (bmb + f) in the framework of the ReLiMo project (Grant No. 0311619). We thank all partners in this project for a fruitful and successful collaboration.

References

1. Rubin, G. M., Yandell, M. D., Wortman, J. R., Miklos, G. L. G., Nelson, C. R., Hariharan, I. K. *et al.* (2000). Comparative Genomics of the Eukaryotes. *Science*, **287**, 2204–2215.
2. Broder, S. & Venter, J. C. (2000). Sequencing the entire genomes of free-living organisms: the foundation of pharmacology in the new millennium. *Annu. Rev. Pharmacol. Toxicol.*, **40**, 97–132.
3. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J. *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
4. Marcotte, E. M., Pellegrini, M., Ng, H., Rice, D. W., Yeates, T. O. & Eisenberg, D. (1999). Detecting protein function and protein–protein interactions from genome sequences. *Science*, **285**, 751–753.
5. Lottspeich, F. (1999). Proteome analysis: a pathway to the functional analysis of proteins. *Angew. Chem. Int. Ed.* **38**, 2476–2492.
6. Wang, J. H. & Hewick, R. M. (1999). Proteomics in drug discovery. *Drug. Discov. Today*, **4**, 129–383.
7. Danchin, A. (1999). From protein sequence to function. *Curr. Opin. Struct. Biol.* **9**, 363–367.
8. Orengo, C. A., Todd, A. E. & Thornton, J. M. (1999). From protein structure to function. *Curr. Opin. Struct. Biol.* **9**, 374–382.
9. Westhead, D. R. & Thornton, J. M. (1998). Protein structure prediction. *Curr. Opin. Biotechnol.* **9**, 383–389.
10. Blundell, T., Jhoti, H. & Abell, C. (2002). High-throughput crystallography for lead discovery in drug design. *Nature Rev. Drug Discov.* **1**, 45–54.
11. Burley, S. K., Almo, S. C., Bonanno, J. B., Capel, M., Chance, M. R., Gaasterland, T., Lin, D., Sali, A.,

- Studier, F. W. & Swaminathan, S. (1999). Structural genomics: beyond the Human Genome Project. *Nature Genet.* **23**, 151–157.
12. Rost, B. (1998). Marrying structure and genomics. *Structure*, **6**, 259–263.
 13. Thornton, J. M., Todd, A. E., Milburn, D., Borkakoti, N. & Orengo, C. A. (2000). From structure to function: approaches and limitations. *Nature Struct. Biol.* **7**, 991–994.
 14. Laskowski, R. A., Luscombe, N. M., Swindells, M. B. & Thornton, J. M. (1996). Protein clefts in molecular recognition and function. *Protein Sci.* **5**, 2438–2452.
 15. Needleman, S. B. & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453.
 16. Waterman, M. S. (1984). General methods for sequence comparison. *Bull. Math. Biol.* **46**, 473–500.
 17. Taylor, W. R. & Orengo, C. A. (1989). Protein structure alignment. *J. Mol. Biol.* **208**, 1–22.
 18. Pearson, W. R. & Lipman, D. J. (1988). Improved Tools for Biological Sequence Analysis. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
 19. Pearson, W. R. & Lipman, D. J. (1990). Rapid and Sensitive Sequence Comparison with FASTP and FASTA. *Methods Enzymol.*, **183**, 63–98.
 20. Bleasby, A. J., Akrigg, D. & Attwood, T. K. (1994). OWL: A non-redundant, composite protein sequence database. *Nucl. Acids Res.* **22**, 3574–3577.
 21. Bairoch, A. & Boeckmann, B. (1994). The S.W.I.S.S.-P.R.O.T. protein sequence databank: current status. *Nucl. Acids Res.* **22**, 3578–3580.
 22. Lessel, U. & Schomburg, D. (1994). Similarities between protein 3D structures. *Protein Eng.* **7**, 1175–1187.
 23. Holm, L. & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123–138.
 24. Grindley, H. M., Artymiuk, P. J., Rice, D. W. & Willett, P. (1993). Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *J. Mol. Biol.* **229**, 707–721.
 25. Nussinov, R. & Wolfson, H. J. (1991). Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. *Proc. Natl Acad. Sci. USA*, **88**, 10495–10499.
 26. May, A. C. W. & Johnson, M. S. (1995). Improved genetic algorithm-based protein structure comparison. *Protein Eng.* **8**, 873–882.
 27. May, A. C. W. & Johnson, M. S. (1994). Protein structure comparison using a combination of a genetic algorithm, dynamic programming and least-squares minimization. *Protein Eng.* **7**, 475–485.
 28. Lehtonen, J. V., Dennessiouk, K., May, A. C. W. & Johnson, M. S. (1999). Finding local structural similarities among families of unrelated protein structures: a generic non-linear alignment algorithm. *Proteins: Struct. Funct. Genet.*, **34**, 341–355.
 29. Richards, F. M. & Kundrot, C. E. (1988). Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. *Protein: Struct. Funct. Genet.* **3**, 71–84.
 30. Mitchell, E. M., Artymiuk, P. J., Rice, D. W. & Willett, P. (1989). Use of graph theory to compare secondary structure motifs in proteins. *J. Mol. Biol.* **212**, 151–166.
 31. Sali, A. & Blundell, T. L. (1990). Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J. Mol. Biol.* **4**, 403–428.
 32. Vriend, G. & Sander, C. (1991). Detection of common three-dimensional substructures in proteins. *Proteins: Struct. Funct. Genet.*, **11**, 52–58.
 33. Alexandrov, N. N., Takahashi, K. & Go, N. (1992). Common spatial arrangements of backbone fragments in homologous and non-homologous proteins. *J. Mol. Biol.* **225**, 5–9.
 34. Pennec, X. & Ayache, N. (1998). A geometric algorithm to find small but highly similar 3D substructures in proteins. *Bioinformatics*, **14**, 516–522.
 35. Wallace, A. C., Borkakoti, N. & Thornton, J. M. (1997). TESS: A geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci.* **6**, 2308–2323.
 36. Artymiuk, P. J., Poirrette, A. R., Grindley, H. M., Rice, D. W. & Willett, P. (1994). A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. *J. Mol. Biol.* **243**, 327–344.
 37. Russell, R. B. (1998). Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J. Mol. Biol.* **279**, 1211–1227.
 38. Kobayashi, N. & Go, N. (1997). A method to search for similar protein local structures at ligand-binding sites and its application to adenin recognition. *Eur. Biophys. J.* **26**, 135–144.
 39. Moodie, S. L., Mitchell, J. B. & Thornton, J. M. (1996). Protein recognition of adenylate: an example of a fuzzy recognition template. *J. Mol. Biol.* **263**, 486–500.
 40. Stahl, M., Taroni, C. & Schneider, G. (2000). Mapping protein surface cavities and prediction of enzyme class by a self-organizing neuronal network. *Protein Eng.* **13**, 83–88.
 41. Fischer, D., Wolfson, H., Lin, S. L. & Nussinov, R. (1994). Three-dimensional, sequence order-independent structural comparison of a serine protease against the crystallographic database reveals active site similarities: potential implications to evolution and to protein folding. *Protein Sci.* **3**, 769–778.
 42. Rosen, M., Liang, S. L., Wolfson, H. & Nussinov, R. (1998). Molecular shape comparisons in searches for active sites and functional similarity. *J. Mol. Biol.* **11**, 263–277.
 43. Laskowski, R. A. (1995). SURFNET: A program for visualizing surfaces, cavities and intermolecular interactions. *J. Mol. Graph.* **31**, 2735–2748.
 44. Lin, S. L., Nussinov, R., Fischer, D. & Wolfson, H. J. (1994). Molecular surface representations by sparse critical points. *Proteins: Struct. Funct. Genet.*, **18**, 94–101.
 45. Klebe, G. (1994). The use of composite crystal-field environments in molecular recognition and the *de novo* design of protein ligands. *J. Mol. Biol.* **237**, 212–235.
 46. Klebe, G. (1993). Structural alignment of molecules. In *3D QSAR and Drug Design: Theory, Methods and Applications* (Kubinyi, H., ed.), pp. 173–199, ESCOM, Leiden.
 47. Hemm, K., Aberer, K. & Hendlich, M. (1995). Constituting a receptor-ligand information base from quality-enriched data. *Ismb*. **3**, 170–178.

48. Hendlich, M. (1998). Databases for protein-ligand complexes. *Acta Crystallogr. D Biol. Crystallog.* **54** (Pt 6), 1178–1182.
49. Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R. & Ferrin, T. E. (1982). A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **161**, 269–288.
50. Levitt, D. G. & Banaszak, L. J. (1992). POCKET: A computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J. Mol. Graph.* **10**, 229–234.
51. Hendlich, M., Rippmann, F. & Barnickel, G. (1997). LIGSITE: Automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph. Model.* **15**, 359–363.
52. Brady, G. P. J. & Stouten, P. F. (2000). Fast prediction and visualization of protein binding pockets with PASS. *J. Comput. Aided Mol. Des.* **14**, 383–401.
53. Liang, J., Edelsbrunner, H. & Woodward, C. (1998). Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci.* **7**, 1884–1897.
54. Edelsbrunner, H. & Mücke, E. P. (1994). Three-dimensional alpha shapes. *ACM Trans. Graph.* **13**, 43–72.
55. Edelsbrunner, H. (1995). The union of balls and its dual shape. *Discrete Comput. Geom.* **13**, 415–440.
56. Peters, K. P., Fauck, J. & Frömmel, C. (1996). The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *J. Mol. Biol.* **256**, 201–213.
57. Hendlich, M., Bergner, A., Günther, J. & Klebe, G. (2002). Relibase—design and development of a database for comprehensive analysis of protein-ligand interactions. *J. Mol. Biol.*, 00 submitted.
58. Günther, J., Bergner, A., Hendlich, M. & Klebe, G. (2002). Utilising structural knowledge in drug design strategies—applications using Relibase. *J. Mol. Biol.*, 00 submitted.
59. Bruno, I. J., Cole, J. C., Lommerse, J. P., Rowland, R. S., Taylor, R. & Verdonk, M. L. (1997). IsoStar: A library of information about nonbonded interactions. *J. Comput.-Aided Mol. Des.*, **11**, 525–537.
60. Bron, C. & Kerbosch, J. (1973). Algorithm 457. Finding all cliques of an undirected graph. *Commun. ACM*, **16**, 575–577.
61. Brint, A. T. & Willett, P. (1987). Algorithms for the identification of three-dimensional maximal common substructures. *J. Chem. Inf. Comput. Sci.* **27**, 152–158.
62. Brint, A. T. & Willett, P. (1989). Upperbound procedures for the identification of similar three-dimensional chemical structures. *J. Comput.-Aided Mol. Design*, **2**, 311–320.
63. Sayle, R. A. & Milner-White, E. J. (1995). RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.* **20**, 374.
64. Fischer, D., Tsai, C. J., Nussinov, R. & Wolfson, H. (1995). A 3D sequence-independent representation of the protein data bank. *Protein Eng.* **8**, 981–997.
65. Hobom, U., Scharf, M., Schneider, R. & Sander, C. (1992). Selection of a representative set of structures from the Brookhaven Protein Data Bank. *Protein Sci.* **1**, 409–417.
66. Hobom, U. & Sander, C. (1994). Enlarged representative set of protein structures. *Protein Sci.* **3**, 522.
67. Strater, N., Schnappauf, G., Braus, G. & Lipscomb, W. N. (1997). Mechanisms of catalysis and allosteric regulation of yeast chorismate mutase from crystal structures. *Structure*, **5**, 1437–1452.
68. Lee, A. Y., Karplus, P. A., Ganem, B. & Clardy, J. (1995). Atomic structure of the buried catalytic pocket of *Escherichia coli* chorismate mutase. *J. Am. Chem. Soc.* **117**, 3327–3362.
69. Hol, W. & Wierenga, R. K. (1984). The α -helix dipole and the binding of phosphate groups of coenzymes and substrates by proteins. In *X-ray Crystallography and Drug Design* (Horn, A. S. a. D. R., C. J., ed.), ???, Oxford.
70. Kinoshita, K., Sadanami, K., Kidera, A. & Go, N. (1999). Structural motif of phosphate-binding site common to various protein superfamilies: All-against-all structural comparison of protein-mono-nucleotide complexes. *Protein Eng.* **12**, 11–14.
71. Gulick, A. M., Song, H., Endow, S. A. & Rayment, I. (1998). X-ray crystal structure of the yeast Kar3 motor domain complexed with Mg. ADP to 2.3 Å resolution. *Biochemistry*, **37**, 1769–1776.
72. Muller-Dieckmann, H. J. & Schulz, G. E. (1995). Substrate specificity and assembly of the catalytic center derived from two structures of ligated uridylate kinase. *J. Mol. Biol.* **246**, 522–530.
73. Bossemeyer, D., Engh, R. A., Kinzel, V., Ponstingl, H. & Huber, R. (1993). Phosphotransferase and substrate binding mechanism of the cAMP-dependent protein kinase catalytic subunit from porcine heart as deduced from the 2.0 Å structure of the complex with Mn²⁺ + adenylyl imidodiphosphate and inhibitor peptide PKI(5-24). *EMBO J.* **12**, 849–859.
74. Zhang, Y., Bond, C. S., Bailey, S., Cunningham, M. L., Fairlamb, A. H. & Hunter, W. N. (1996). The crystal structure of trypanothione reductase from the human pathogen *Trypanosoma cruzi* at 2.3 Å resolution. *Protein Sci.* **5**, 52–61.
75. Tanaka, N., Nonaka, T., Nakanishi, M., Deyashiki, Y., Hara, A. & Mitsui, Y. (1996). Crystal structure of the ternary complex of mouse lung carbonyl reductase at 1.8 Å resolution: the structural origin of coenzyme specificity in the short-chain dehydrogenase/reductase family. *Structure*, **4**, 33–45.
76. Schluckebier, G., Kozak, M., Bleimling, N., Weinhold, E. & Saenger, W. (1997). Differential binding of S-adenosylmethionine S-adenosylhomocysteine and Sinefungin to the adenine-specific DNA methyltransferase M.TaqI. *J. Mol. Biol.* **265**, 56–67.
77. Enroth, C., Neujahr, H., Schneider, G. & Lindqvist, Y. (1998). The crystal structure of phenol hydroxylase in complex with FAD and phenol provides evidence for a concerted conformational change in the enzyme and its cofactor during catalysis. *Structure*, **6**, 605–617.
78. Breton, R., Housset, D., Mazza, C. & Fontecilla-Camps, J. C. (1996). The structure of a complex of human 17 β -hydroxysteroid dehydrogenase with estradiol and NADP + identifies two principal targets for the design of inhibitors. *Structure*, **4**, 905–915.
79. Marquart, M., Walter, J., Deisenhofer, J., Bode, W. & Huber, R. (1983). The geometry of the reactive site and the peptide groups in trypsin, trypsinogen and its complexes with inhibitors. *Acta Crystallogr.* **B39**, 480–490.
80. Gallagher, D. T., Pan, Q. W. & Gilliland, G. L. (1998). Mechanism of ionic strength dependence of crystal growth rates in a subtilisin variant. *J. Cryst. Growth*, **193**, 665–673.

81. Cho, H. S., Ha, N. C., Choi, G., Kim, H. J., Lee, D., Oh, K. S. *et al.* (1999). Crystal structure of delta(5)-3-ketosteroid isomerase from *Pseudomonas testosteroni* in complex with equilenin settles the correct hydrogen bonding scheme for transition state stabilization. *J. Biol. Chem.* **274**, 32863–32868.
82. Almog, O., Gallagher, T., Tordova, M., Hoskins, J., Bryan, P. & Gilliland, G. L. (1998). Crystal structure of calcium-independent subtilisin BPN' with restored thermal stability folded without the pro-domain. *Proteins: Struct. Funct. Genet.*, **31**, 21–32.
83. Martin, J. L., Begun, J., Schindeler, A., Wickramasinghe, W. A., Alewood, D., Alewood, P. F. *et al.* (1999). Molecular recognition of macrocyclic peptidomimetic inhibitors by HIV-1 protease. *Biochemistry*, **38**, 7978–7988.
84. Narayana, N., Diller, T. C., Koide, K., Bunnage, M. E., Nicolaou, K. C., Brunton, L. L. *et al.* (1999). Crystal structure of the potent natural product inhibitor balanol in complex with the catalytic subunit of cAMP-dependent protein kinase. *Biochemistry*, **38**, 2367–2376.
85. Tyndall, J. D., Reid, R. C., Tyssen, D. P., Jardine, D. K., Todd, B., Passmore, M. *et al.* (2000). Synthesis, stability, antiviral activity, and protease-bound structures of substrate-mimicking constrained macrocyclic

inhibitors of HIV-1 protease. *J. Med. Chem.*, **43**, 3495–3504.

Edited by R Huber

(Received 20 March 2002; received in revised form 10 July 2002; accepted 17 July 2002)



<http://www.academicpress.com/jmb>

Supplementary Material for this paper comprising three tables listing the best ranked solutions for the comparisons of a NADPH binding pocket (1cyd), a trypsin cavity (1tpo) and a subtilisin cavity (1sua) are available on IDEAL