doi:10.1016/j.jmb.2006.04.024 *J. Mol. Biol.* (2006) **359**, 1023–1044

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

From the Similarity Analysis of Protein Cavities to the Functional Classification of Protein Families Using Cavbase

Daniel Kuhn[1], Nils Weskamp[1,2], Stefan Schmitt[3], Eyke Hüllermeier[2] and Gerhard Klebe[1]*

[1]*Department of Pharmaceutical Chemistry, University of Marburg, Marbacher Weg 6 D-35032 Marburg, Germany*

[2]*Department of Mathematics and Computer Science University of Marburg D-35032 Marburg, Germany*

[3]*AstraZeneca R&D, S-43183 Mölndal, Sweden*

**Corresponding author*

In this contribution, the classification of protein binding sites using the physicochemical properties exposed to their pockets is presented. We recently introduced Cavbase, a method for describing and comparing protein binding pockets on the basis of the geometrical and physicochemical properties of their active sites. Here, we present algorithmic and methodological enhancements in the Cavbase property description and in the cavity comparison step. We give examples of the Cavbase similarity analysis detecting pronounced similarities in the binding sites of proteins unrelated in sequence. A similarity search using SARS M^{pro} protease subpockets as queries retrieved ligands and ligand fragments accommodated in a physicochemical environment similar to that of the query. This allowed the characterization of the protease recognition pockets and the identification of molecular building blocks that can be incorporated into novel antiviral compounds. A cluster analysis procedure for the functional classification of binding pockets was implemented and calibrated using a diverse set of enzyme binding sites. Two relevant protein families, the α-carbonic anhydrases and the protein kinases, are used to demonstrate the scope of our cluster approach. We propose a relevant classification of both protein families, on the basis of the binding motifs in their active sites. The classification provides a new perspective on functional properties across a protein family and is able to highlight features important for potency and selectivity. Furthermore, this information can be used to identify possible cross-reactivities among proteins due to similarities in their binding sites.

Keywords: protein binding pockets; classification of protein binding pockets; cluster analysis of protein binding pockets; protein kinases; SARS protease

Introduction

Protein function, in particular that of enzymes, is often intimately connected with the recognition and chemical modification of endogenous ligands, such as agonists, antagonists, effectors and substrates. This recognition usually occurs in well-characterized cavities or binding sites on the protein surface. Due to the functional importance of protein binding sites, there are several important applications of their similarity analysis: (i) several studies have demonstrated that there is not necessarily a correlation between the fold and function of proteins.[1–4] Accordingly, similarity analysis of binding sites can complement methods based only on information about sequence and fold in the

Present address: D. Kuhn, Boehringer Ingelheim Austria, Department of Medicinal Chemistry, A-1121 Vienna, Austria.

Abbreviations used: SAM, *S*-adenosyl-methionine; FAD, flavine adenine dinucleotide; SARS, severe acute respiratory syndrome; CoV, coronavirus; TGEV, transmissible gastroenteritis virus; CA, carbonic anhydrase; CML, chronic myelogenous leukemia; MAP, mitogen-activated protein kinases; CDKs, cyclin-dependent protein kinases; HB, hydrogen bond; RMSD, root-mean-square deviations; UPGMA, unweighted pair group method with arithmetic mean; EC, enzyme classification.

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

functional annotation of protein structures. (ii) With respect to the development of novel leads, the knowledge about ligands and ligand fragments that bind to structurally and physicochemically similar (sub-) pockets in other proteins can provide important ideas for drug discovery, in particular with respect to *de novo* design or the bioisosteric replacements of molecular building blocks. (iii) We showed recently that binding site similarities detected between unrelated proteins can help to rationalize and predict cross-reactivities.[5] Different proteins with similarities in their binding site recognition properties may bind the same drugs. (iv) Furthermore, a classification of proteins based on the similarities between their active sites can be derived. This alternative taxonomy provides another perspective on the protein space and possibly highlights relationships between proteins, including conformational adaptations, that might not be apparent using the well-established comparative tools.

There are several methods published in the literature that compare protein structures and detect common patterns between them. These methods can be divided into two categories: methods based on predefined common three-dimensional templates comprised of several amino acids,[6–10] and methods that operate independently of any reference template, mutually comparing entire protein structures or predefined regions of interest.[11–31] Template-based methods usually represent the amino acids by several pseudo-atoms that encode the properties of the amino acids. Based on these descriptors, they are able to retrieve protein structures exhibiting similar patterns. However, the search results will be biased, to some extent, by the choice of predefined query template. Methods that use the entire protein structure or a predefined portion can range from simple (e.g. C^{α} coordinates) to rather complex descriptors such as Connolly surface points containing geometric information (surface shape descriptors) or physicochemical information (amino acid type, hydrophobicity, electrostatic potential). To identify common substructures between assigned descriptors, different algorithms such as geometric hashing,[8,11–22] genetic algorithms,[27,28] graph matching algorithms,[6,7,25,26,29–31] string matching,[23,24] and searches in multi-index structures[10] have been used.

We recently introduced Cavbase, a method for describing and comparing protein binding pockets.[32] Binding pockets on the protein surface are detected automatically, and their physicochemical properties are encoded by five generic descriptors (assigned pseudocenters). A clique algorithm detects similar arrangements of pseudocenters between two cavities. Here, we present the optimization of the physicochemical representation of a binding pocket validated against experimental data using an improved set of pseudocenter definitions, we retrieved further examples showing structurally and functionally related cavities independent of fold and sequence homology. A similarity search based on the individual subpockets of SARS M^{pro} protease is used to demonstrate the scope of Cavbase to characterize protein binding pockets with respect to bound ligands and ligand fragments. Once appropriate ligand fragments are detected in subpockets of similar proteins, their geometry may serve as guidelines for further inhibitor design. All of the above-mentioned methods for comparing protein structures focus on the comparison of selected binding pockets or query patterns against a database of protein binding sites. Here, we extend the mutual similarity analysis by clustering a sample of protein cavities to discover relationships between binding sites of protein families. Such criteria will be used to differentiate between protein families and, at the same time, to detect similarities among entries originating from the same protein family. To achieve such a classification, a clustering procedure was implemented and optimized. Finally, the classification of two pharmaceutically relevant protein families, the α-carbonic anhydrases and the eukaryotic protein kinases, in terms of the Cavbase taxonomy are presented.

Results and Discussion

Optimization of Cavbase property description

To assess the quality by which the pseudocenters and surface patches in Cavbase describe the interaction properties exposed to a binding site, the Isostar database was consulted.[33] In this database, crystallographic data from the CSD[34] and PDB[35] have been collected in terms of scatter plots containing the distribution of certain contact groups around a central group of interest, e.g. a functional group found in an amino acid. This information was evaluated to derive the definitions of the pseudocenter positions, during the design of Cavbase.[32] Since a stringent pseudocenter definition allowing for fast binding site comparisons should be achieved, the properties of some residues were not described in full detail in our first implementation. Therefore, we have reassigned the pseudocenter definitions in a more elaborate fashion. For example, the hydrogen bonding properties of cysteine were not considered; nevertheless, the sulfhydryl group can be involved in hydrogen bonding. McDonald and co-workers found that cysteine rarely serves as hydrogen-bond acceptor but often acts as hydrogen-bond donor.[36] Furthermore, cysteine is often found as a catalytically relevant residue in protein active sites.[37] To account for its hydrogen bonding ability, a donor center was added at the sulfur atom of cysteine (Figure 1). In the case of methionine, there seems to be no significant involvement of the sulfur atom in hydrogen bonding. Accordingly, its side-chain remains represented by one aliphatic pseudocenter only .

The side-chains of asparagine, glutamine, aspartate, glutamate, and arginine are able to form π–π

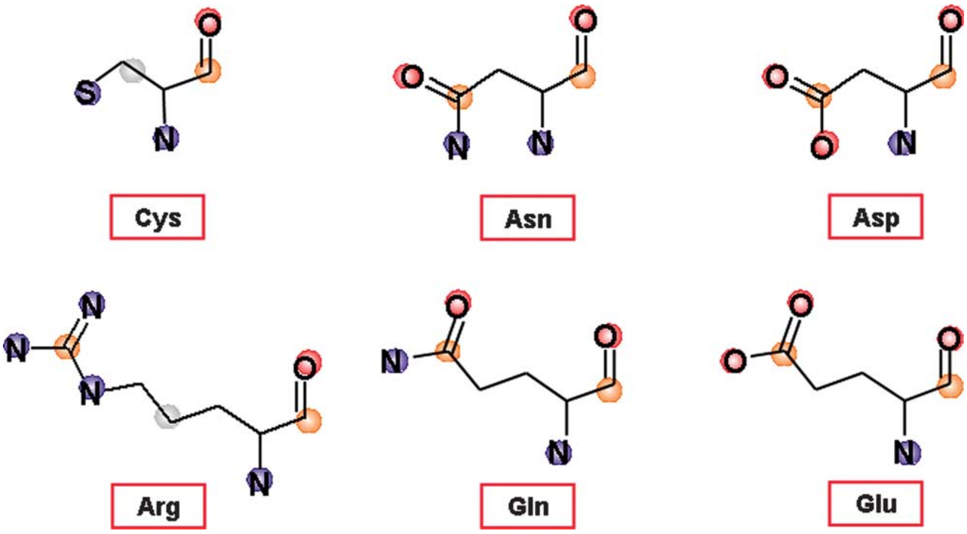


Figure 1. The expanded definition of pseudocenters in Cavbase. A donor pseudocenter (blue sphere) was introduced at the side-chain of cysteine to account for the hydrogen bond donor properties of cysteine. Pi pseudocenters (orange spheres) were introduced at terminal side-chains of asparagine, aspartate, glutamine, glutamate, and arginine, reflecting their ability to form π–π interactions with neighboring functional groups, including those of ligands. Hydrogen bond acceptor and aliphatic pseudocenters are shown as red and white spheres, respectively.

interactions with neighboring functional groups, including those of ligands. Such interaction properties are evident from the analysis of amino acids exhibiting aromatic moieties,[38,39] or from consulting Isostar. The distribution of aromatic contact groups around terminal amide groups shows a clear preference for interactions perpendicular to a best plane defined by the atoms of the amide group. To represent these interaction properties, a pi pseudocenter located at the carbon atom of the carboxylate, carboxamide and guanidine group has been introduced (Figure 1). The ability of carboxylate groups to form π–π interactions is regarded to be weaker than that of carboxamide groups.[38] Nevertheless, the analysis of crystal data gives clear evidence that this type of interaction occurs. As a result, two of the three pseudocenters assigned to the terminal group of e.g. glutamine and glutamate, are recognized as similar (Figure 1).

Overall validation of the Cavbase descriptors using Drugscore

To further validate and assess the quality of the pseudocenter and surface patch description in Cavbase, a comparison with hotspots obtained by Drugscore was performed.[40,41,42] For this purpose, Drugscore pair potentials were selected that correspond to the physicochemical properties of the pseudocenters. Gohlke *et al*. used the atom types listed in Table 1.[40] A set of 214 proteins obtained from CCDC/Astex was used to compare the spatial location of Drugscore hotspots with the Cavbase surface description.[43] The correspondence between the Drugscore hotspots and the analogous cavity surface patches in Cavbase was examined visually. To evaluate the fit, the Drugscore hotspots were projected onto the cavity surface. Good agreement of the two descriptions was found. In Figure 2, the cavity of a dihydro-orotate-dehydrogenase (PDB code 1d3h) is displayed as a dotted surface together with three different Drugscore hotspot regions using three probe atoms (Drugscore atom type N.3 (Figure 2-II), O.2 (Figure 2-III) and C.ar (Figure 2-IV)). It is particularly notable that the hotspots for directional interactions (hydrogen bond donors and acceptors) are in very good agreement with the Cavbase property description, suggesting satisfactory representation by our pseudocenter description.

Representation of the π interaction property of aromatic amino acids

Our initial implementation of Cavbase considered the interaction properties of aromatic moieties insufficiently, because the possible edge-to-face interactions of aromatic moieties in amino acid side-chains were neglected. To assess the accessibility of residues to the cavity surface, the angle enclosed by the two vectors **v** and **r** serves as a criterion to decide whether a particular pseudocenter is considered or discarded. The original cut-off value of 60° assigned to π-interactions of aromatic moieties was set too low to properly account for edge-to-face interactions. Therefore, two cut-off values were introduced, depending on the origin of the pi pseudocenter. In the case of pi pseudocenters originating from an aromatic side-chain, the cut-off value was set to 100°, whereas in the case of the pi pseudocenters originating from peptide bonds, the 60° value still appeared valid. The changes resulting from the readjustment of the

Table 1. Drugscore probe atoms used to analyze the corresponding physicochemical property in Cavbase

Drugscore atom type	Cavbase property	Hydrophobic/hydrophilic	Physicochemical interaction type
C.3	Aliphatic	Hydrophobic	Aliphatic
C.ar	Aromatic	Hydrophobic	Aromatic
O.3	Donor acceptor	Hydrophilic	Hydrogen bond acceptor and donor
O.2	Acceptor	Hydrophilic	Hydrogen bond acceptor
N.3	Donor	Hydrophilic	Hydrogen bond donor

cut-off value are shown, using a hydrolase as an example (Figure 3).

Searching for similar binding sites

Several examples of structural and physicochemical similarities between binding sites of proteins unrelated in sequence are described in the literature.[20,23,28,32,44] The most prominent example of functional similarities between proteins from two distinct fold families is that of the serine proteases trypsin and subtilisin. The two proteins show low levels of sequential and structural homology.[32,45,46] However they catalyze the same chemical reaction, which requires a similar distribution of physicochemical properties in their binding sites. As described in our previous communication on Cavbase,[32] the approach is able to retrieve a cavity of the trypsin family from a large set of diverse binding pockets using a subtilisin cavity as a query, or *vice versa*. In the following, we present further examples of the retrieval of cavities from proteins with similar function but distinct fold.

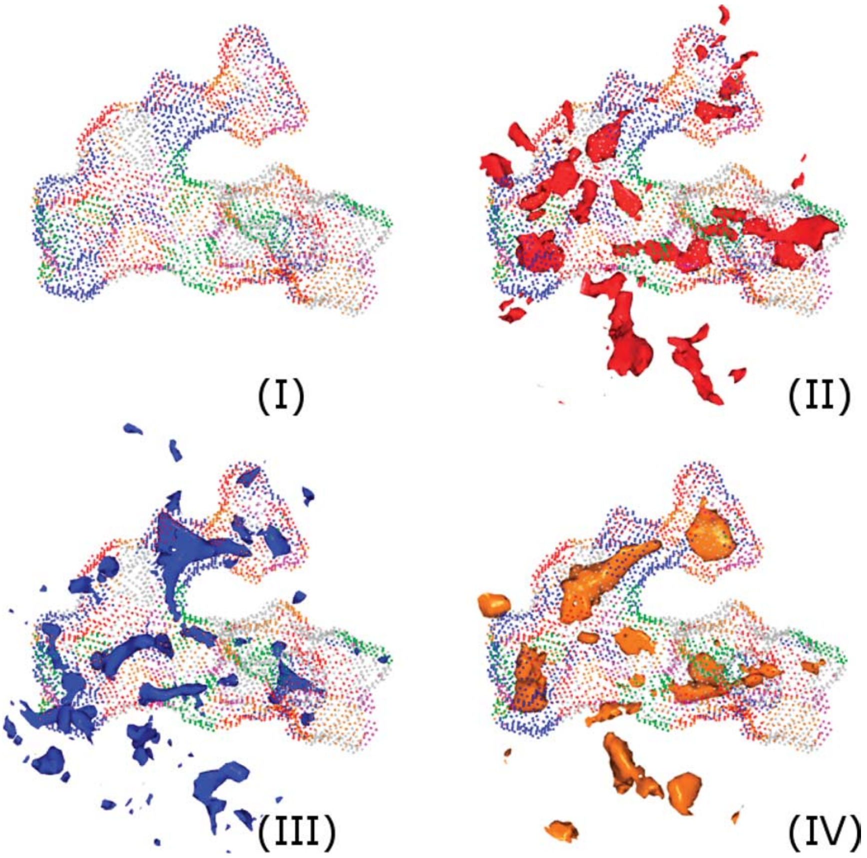


Figure 2. Validation of property surface patches in Cavbase by comparison with Drugscore maps. A comparison of the Drugscore hotspots with the Cavbase binding site description is shown for the binding pocket of dihydro-orotate-dehydrogenase (PDB code 1d3h). In (I) the Cavbase surface patches, annotated with respect to the five physicochemical properties of the neighboring pseudocenters, are shown as dotted surfaces (color scheme used: H bond donor (blue), H bond acceptor (red), ambivalent donor/acceptor (green), hydrophobic aliphatic (white) or aromatic/pi (orange)). (II) to (IV) display three types of Drugscore hotspots, together with the Cavbase surface. The color coding of the hotspots corresponds to that of the related Cavbase properties (Drugscore atom type N.3 (red, II), O.2 (blue, III) and C.ar (orange, IV)). Drugscore hotspots that describe directional interactions match very well with the corresponding Cavbase surface patches. The contour levels are calibrated for each atom type in such a way that 0.6% of the grid points are assigned to the most favorable interaction areas.

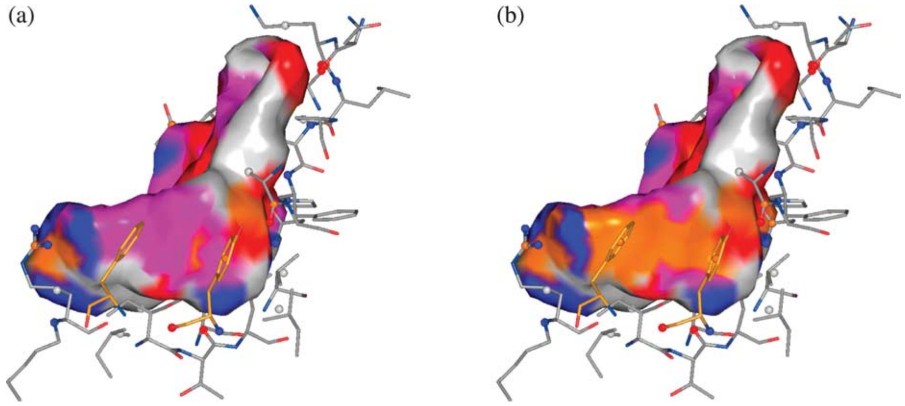


Figure 3. Consideration of edge-to-face interactions of aromatic moieties in the Cavbase binding site description and similarity searches. The binding pocket of a hydrolase (PDB code 1tum) is used to demonstrate the influence of different parameter settings for pseudocenters describing hydrophobic aromatic interactions, with respect to their exposure onto the protein surface (color coding as in Figure 2). Two phenylalanine residues (carbon atoms colored orange) can potentially perform edge-to-face interactions towards the cavity surface. Using the original angular parameter settings (left), no interaction towards the cavity surface would be considered for either ring (violet areas). Recalibrating the angle between the standard vector **r** and the mean orientation vector **v** to 100° (right) allows the recognition of edge-to-face interactions, and both pi pseudocenters can now expose their property onto the cavity surface (orange areas).

NAD(P)-binding enzymes

The search for NAD(P) sites in a diverse set of proteins is presented as an additional example of the retrieval of binding pockets that bind the same cofactor for a chemical reaction, but for which the binding involves different amino acids. The binding site in UDP-galactose-4-epimerase (PDB code 1xel)[47] was compared against dataset I (see Materials and Methods). It is relatively large and encompasses, besides NADH, a binding site for UDP-galactose. The epimerase catalyzes the conversion of UDP-galactose to UDP-glucose, simultaneously reducing NAD^+. The product UDP-glucose is bound together with NADH in the query structure. As expected, binding sites hosting NADH were found amongst the most highly scored cavities. The similarity across the entire cofactor-binding site is quite pronounced, since the same amino acids are involved in cofactor coordination and little structural variation is observed. As an example, Figure 4 depicts the areas detected as similar in the binding sites of the UDP-galactose-4-epimerase and acyl-CoA-dehydrogenase (PDB code 1e6w). At subsequent positions in the cavity ranking, binding pockets that accommodate related cofactors such as *S*-adenosyl-methionine (SAM) and flavine adenine dinucleotide (FAD) are detected.

Interestingly, the nucleotide-binding site of a glucose oxidase (PDB code 1gal) is placed at rank 149 with bound FAD. The glucose oxidase shows no sequence identity (19.7%) or structural fold similarity to the query protein; however, the two proteins possess similar sequence motifs, which are involved in the coordination of the phosphate backbone of the nucleotide.[48] The UDP-galactose-4-epimerase adopts a Rossmann fold, whereas the glucose oxidase exhibits an FAD/NAD(P)-binding domain fold. The amino acids involved directly in cofactor binding are entirely different; nevertheless, they expose similar physicochemical properties to the cofactor binding site, successfully recognized by Cavbase (Table 2). Figure 4 depicts areas detected as similar in the two binding cavities.

SARS-coronavirus M^{pro}

The severe acute respiratory syndrome (SARS) is an atypical pneumonia that originated in Southern China and spread over 30 countries in the first half of 2003. It is caused by a novel coronavirus (CoV), the SARS CoV.[49,50] The virus secretes the protease M^{pro} (3C-like, 3CL), which exhibits an important function in the viral life-cycle.[51] Its inhibition provides a promising therapeutic principle for the development of anti-viral drugs against SARS. The folding pattern of the CoV M^{pro} is related to the fold of serine proteases (chymotrypsin type),[52,53] which consist of two domains (I and II). Additionally, the SARS CoV M^{pro} exhibits a third helical domain. The binding pocket at the interface between domain I and domain II, comprising the residues of the catalytic dyad (His45 and Cys145), is detected by Ligsite. This pocket has been used for a subsequent similarity search with Cavbase against dataset I. In the reference crystal structure of SARS-CoV, a peptidic inhibitor (**1**) (PDB code 1uk4)[54] is bound. Cavbase retrieves the human homologues of SARS CoV M^{pro}, which are most highly ranked, followed by other viral proteases such as human CoV, transmissible gastroenteritis virus (TGEV) and the tobacco etch virus. Interestingly, the rhinovirus 3C protease with the bound inhibitor AG7088 (**2**) is found at rank 13.[55,56] In Figure 5, the superposition of the SARS CoV M^{pro} and the latter rhinoviral protease is shown. The residues comprising the catalytic site and the main-chain substrate recognition site are detected as similar in the two pockets. Analogous side-chains of both inhibitors address

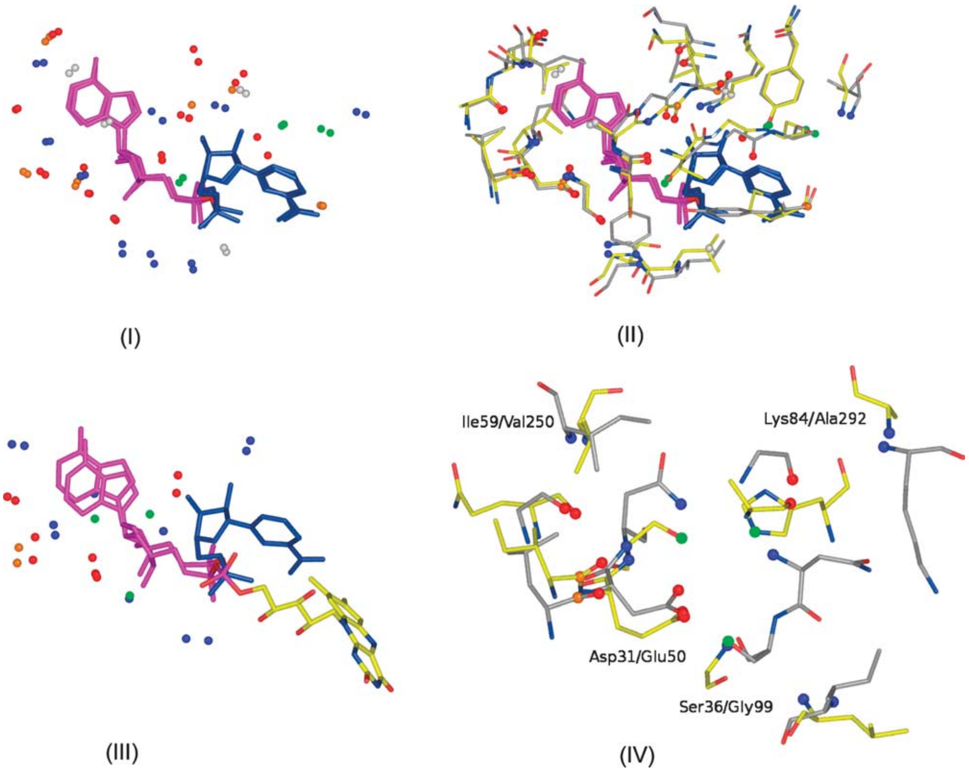


Figure 4. The nucleotide cofactor binding sites in UDP-galactose-4-epimerase (PDB code 1xel, grey carbon atoms) and acyl-CoA-dehydrogenase (PDB code 1e6w, yellow carbon atoms) detected as similar by Cavbase. The NADH binding is performed by virtually identical amino acids. In (I), the matching pseudocenters (color coding as in Figure 2) in both binding pockets are shown, together with NADH. In (II), the matching amino acids are displayed in addition, suggesting pronounced structural conservation between the two binding sites. The comparison of the UDP galactose with a glucose oxidase (PDB code 1gal, yellow carbon atoms) reveals a case where the nucleotide cofactor binding is performed by entirely different amino acids, although their physicochemical interactions are similar, and this is detected by Cavbase. The matching pseudocenters and cofactors (NADH (1xel) and FAD (1gal)) in the two binding sites are shown in (III). In (IV), the amino acids superimposed on the matched pseudocenters, are displayed.

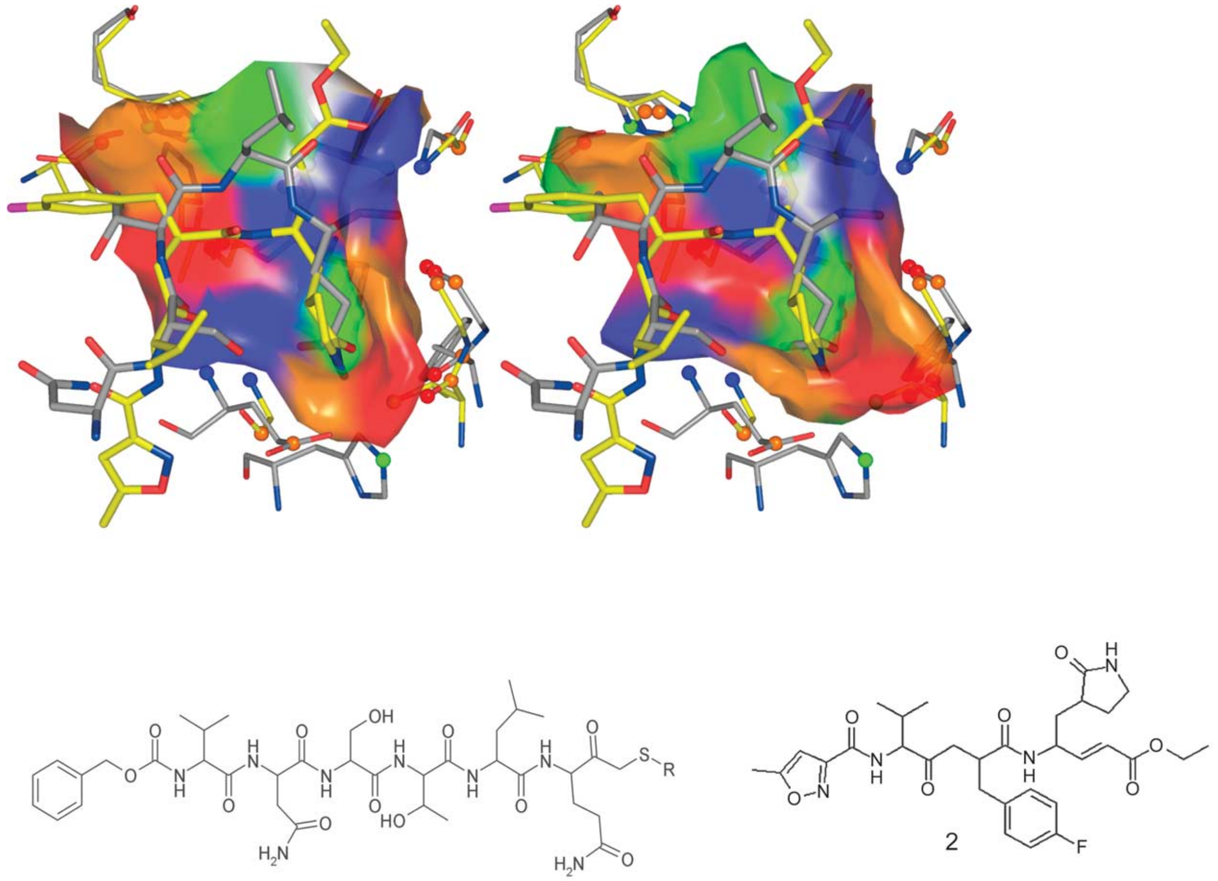
corresponding subpockets in the proteases, and matching pseudocenters are superimposed convincingly (RMSD = 1.046 Å). Even though the Cavbase approach does not consider any coordinate of bound inhibitors, the resulting alignment indicates analogous side-chains of the two inhibitors addressing corresponding subpockets in the proteases.

Over the following 200 ranks, cavities from members of the serine protease family are frequently found. The binding pocket of an α-lytic protease (PDB code 6lpr) was the most highly

Table 2. Matched pseudocenters and amino acids in the binding sites of UDP-galactose-4-epimerase and glucose oxidase

UDP galactose-4-EPimease (1xel)			Glucose oxidase (1gal)		
Pseudocenter type	Corresponding amino acid		Pseudocenter type	Corresponding amino acid	
Donor	Ile12	p	Donor	Leu29	p
Pi	Leu30	p	Pi	Ile49	p
Acceptor	Leu30	p	Acceptor	Ile49	p
Acceptor	Asp31	s	Acceptor	Glu50	s
Acceptor	Asp31	s	Acceptor	Glu50	s
Donor	Asn32	p	Donor	Ser51	p
Donor	Asn32	s	Donor-acceptor	Ser51	s
Donor	Asn35	p	Donor-acceptor	His78	s
Donor-acceptor	Ser36	s	Donor	Gly99	p
Acceptor	Gly57	p	Acceptor	Gln248	p
Donor	Ile59	p	Donor	Val250	p
Acceptor	Gly82	p	Acceptor	Ala289	p
Donor	Lys84	p	Donor	Ala292	p

The three-letter code is used for the amino acid, the number of the amino acid and the origin of the pseudocenter, from the side-chain (s) or from the peptide bond (p).



1

2

Figure 5. The superposition of the matched protein side-chains, pseudocenters and bound inhibitors from the comparison of SARS M^{pro} (grey carbon atoms) and rhinovirus 3C protease (yellow carbon atoms). The SARS M^{pro} inhibitor peptide **1** and the rhinoviral inhibitor AG7088 **2** superimpose convincingly. On the left, the surface patches found by Cavbase for the SARS protease are shown; on the right, those of the rhinovirus 3C protease are displayed. The surface patches are color-coded according to the corresponding physicochemical properties (see Figure 2).

ranked non-viral protein, at rank 14. The residues of the catalytic dyad/triad, parts of the main-chain recognition region and residues forming the oxyanion hole are matched. Accordingly, the areas responsible for peptide cleavage reaction are matched between the two proteases.

Similarity searching using SARS CoV M^{pro} subpockets

The previous functional comparison was performed using the complete binding site of SARS CoV M^{pro}. To focus the analysis on subpocket properties of SARS CoV M^{pro} (PDB code 1uk4), the various subpockets were dissected and used as individual query cavities (subpocket labeling according to the human CoV M^{pro} structure (PDB code 1p9u)).[52] Once related subpockets were matched in other proteins, the information about ligand fragments bound in these structures can be used to assist in the design of novel antiviral drugs. The pseudocenters describing the properties of the various subpockets were selected visually and their accessibility was verified. The S1 subpocket is described by a set of pseudocenters assigned to the amino acid residues Phe140, Leu141, Ser144, His163, Met165, Glu166, and His172, and the S2 subpocket is described by a set of pseudocenters assigned to the amino acid residues His41, Met49, Met165, Asp187, Arg188, and Gln189 (Figure 6). The S1 subpocket of the SARS protease exhibits a polar character, whereas the S2 subpocket possesses a more aromatic and aliphatic character. The S3 pocket has been ignored, as it is highly solvent-exposed. Since the S4 pocket is partially solvent-exposed and forms a rather shallow subpocket, it is not well suited for a similarity analysis using Cavbase. Accordingly, it has not been considered in the similarity searching. The S1 and S2 subpockets were compared against dataset I.

The first 250 ranked cavities of each similarity search were inspected visually and the bound ligands of the matched binding sites were classified into seven generic groups according to their

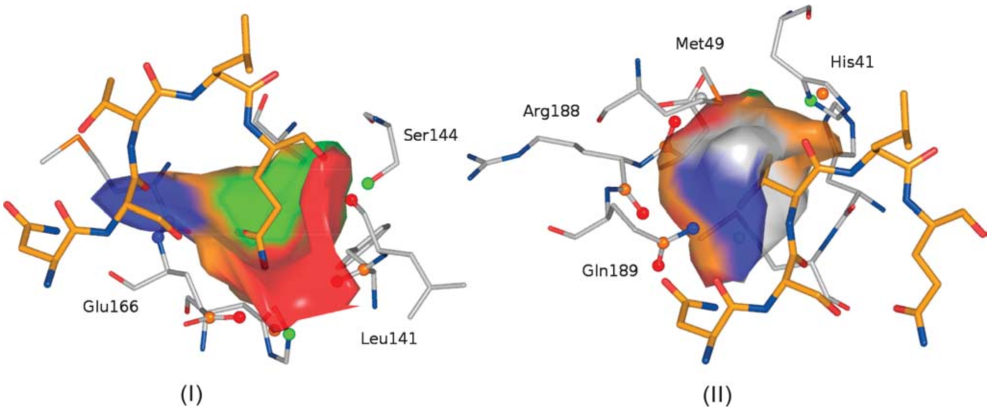


Figure 6. Representation of the SARS CoV M^{pro} S1 (I) and S2 (II) subpockets, with the physicochemical properties of the adjacent pseudocenters mapped onto the surface (color coding as in Figure 2). The considered pseudocenters were selected in such a way as to fully characterize each subpocket. The S1 subpocket exhibits a stronger polar character, whereas the S2 subpocket exhibits an aliphatic-aromatic character.

chemotype (Figure 7): ligand fragments with an alcohol/acid group, with a basic group (e.g. amine, guanidine), with a phosphate group, with a sugar group, with an aliphatic group, with an aromatic group, or with heme groups (Table 3). This strategy allows classification of the subpockets in terms of the physicochemical properties of the ligand fragments that are found in each subpocket.

In case of the S1 subpocket search, 87 binding sites that host a ligand fragment that spatially superimposes on the side-chain of glutamine of the peptidic SARS inhibitor could be identified within the first 250 ranks. The majority of the retrieved ligand fragments exhibit a polar character, represented by sugar or phosphate groups or by ligand moieties with basic character (e.g. amino or guanidino groups). The physicochemical properties of the detected ligands match well the cleavage preference of the SARS CoV M^{pro}. Substrates to be cleaved by CoV M^{pro} have glutamine in this position.[51]

A total of 118 cavities could be identified within the first 250 hits of the S2 subpocket search that contained a ligand fragment that spatially superimposes on the side-chain of the peptidic SARS inhibitor. The analysis of the chemotypes of the retrieved ligands reveals that, in this subpocket, the bound ligands possess predominantly aliphatic and aromatic character, e.g. alkyl chains, phenyl, imidazole, or pyrimidine rings. According to the cleavage preference of the SARS protease, leucine is preferred at this position. In the SARS CoV M^{pro} structure used as a reference, the peptide inhibitor adopts an unusual binding mode. The P3 amino acid residue (threonine) fills the S2 pocket, whereas the P2 residue (leucine) points towards the solvent. Interestingly, in our cavity search 12 ligands composed of serine-like or threonine-like fragments are discovered, and all superimpose convincingly on the threonine residue found at this position in SARS CoV M^{pro}. The additional polar alcohol group seems to be tolerated at this position.

The affinity of **2**, and analogous compounds, towards SARS CoV M^{pro} was determined. Whereas **2** shows no inhibition, closely related analogs of **2** exhibit IC$_{50}$ values in the two-digit micromolar range.[57] Analysis of the fragments that occupy the S2 subpocket in the most active members of this series reveals phenyl and isopropyl groups at this position. Additionally, Yang *et al*. could solve the crystal structures of SARS CoV M^{pro} with several inhibitors based on the skeleton of **2**, with phenylalanine or leucine side-chains filling the corresponding S2 pocket.[58] The results from the similarity analysis performed by Cavbase are in excellent agreement with these subsequently published experimental findings.

We could further utilize the results of the Cavbase similarity search and subpocket characterization in the design of a library of peptide aldehydes.[84] The most potent compounds show IC$_{50}$ values in the one-digit micromolar range.

Classification of protein families

Validation of clustering procedure

A cluster analysis based on Cavbase similarities can be performed using different combinations of clustering algorithms, with appropriately set clustering parameters, and the three Cavbase scoring functions. First, an optimal clustering setup had to be found. A dataset of 105 cavities from 13 functionally diverse enzyme families was used to calibrate the clustering procedure. In Table 4, the PDB codes, the considered protein families and the EC codes, together with the SCOP superfamily and family, are listed‡. All entries from one enzyme family belong to the same SCOP subfamily;

‡ The protein kinases from the serine/threonine (2.7.1.37) and the tyrosine subfamily (2.7.1.112) are regarded as one enzyme family.

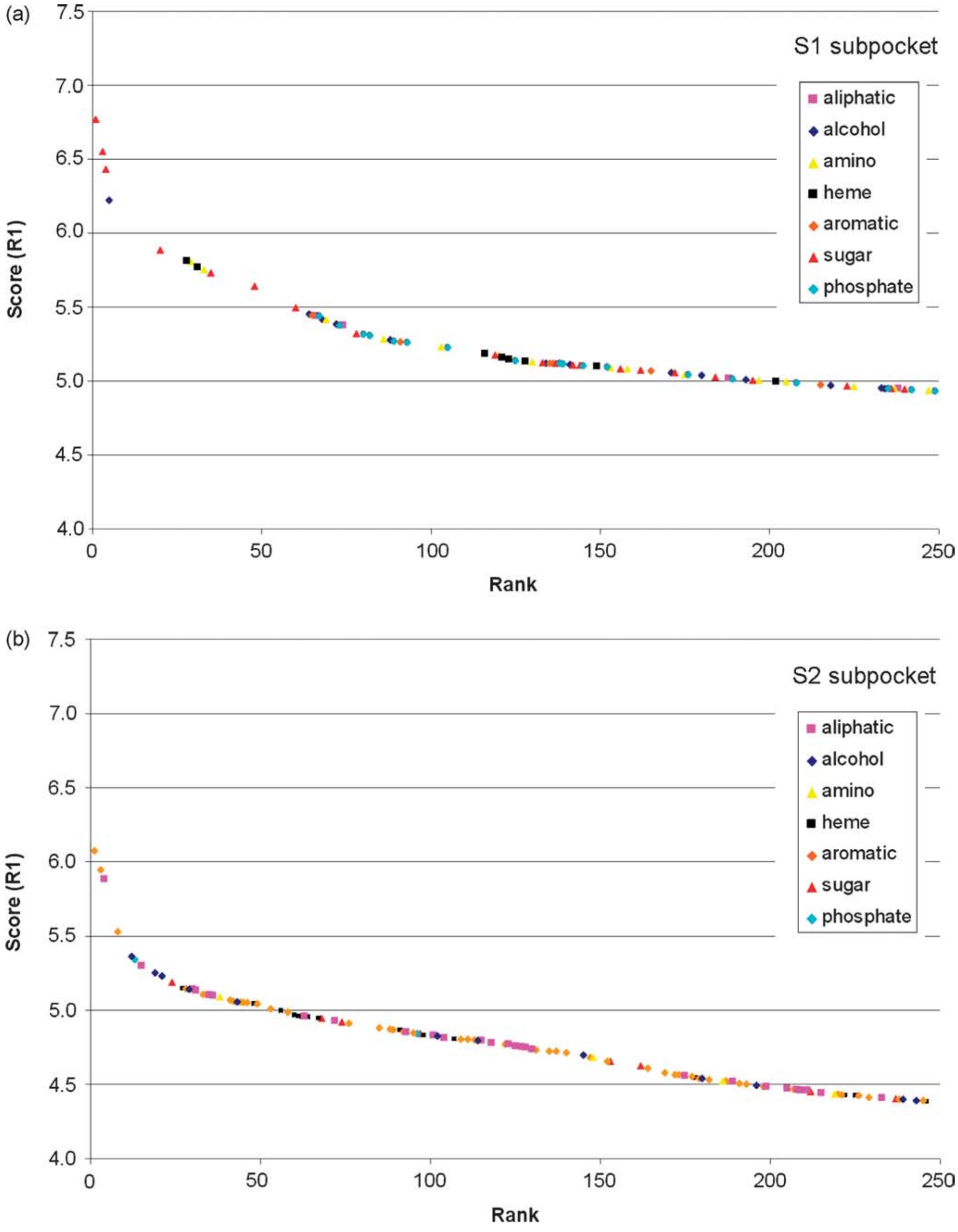


Figure 7. Analysis of the ligands and ligand fragments found in the 250 best ranked binding pockets that could be superimposed with the peptidic SARS inhibitor using the SARS CoV M$^{\mathrm{pro}}$ S1 and S2 subpockets as query cavities in a Cavbase similarity search. The ranking was performed using the R_1 scoring function. The binding pockets are classified according to seven generic chemotypes, by which the bound ligand is characterized (Table 3).

therefore, similarities in sequence and in the folding pattern are found across the family members. Each protein entry was checked carefully to include only the "catalytic" cavity into the dataset; i.e. the cavities comprising the known catalytic residues.

The Cavbase similarity scores do not obey the prerequisites for an optimal similarity metric, which requires reflexivity (cavity A is similar to itself), symmetry (if cavity A is similar to cavity B, then cavity B is similar to cavity A), and transitivity (if cavity A is similar to B and cavity B is similar to cavity C, then cavity A is also similar to cavity C). Therefore, their suitability for a cluster analysis should be investigated. Whereas the first two properties are met, the transitive term is not necessarily met. In total, 12 different combinations of algorithms and scoring functions (four clustering algorithms together with three Cavbase scoring functions) were systematically evaluated as setup for clustering. One crucial parameter in clustering is the definition of the number of predefined output clusters. This value defines the level of detail to which the clustering process will analyze the data. A low number of clusters will merge dissimilar cavities into one cluster, whereas too large a number will likely obscure the detection of patterns in a dataset, and lead to the generation of many singletons. In principle, any completed hierarchical

Table 3. Statistics on the chemotype of the ligands and ligand fragments found in highly scored binding pockets, taking the S1 and S2 subpockets of SARS CoV M^{pro} as queries

	Occurrence	
Bound ligand chemotype	S1 subpocket	S2 subpocket
Ligands with basic groups (amines)	16	4
Aliphatic ligands	4	28
Ligands with phosphate groups	19	2
Ligands with sugar groups	21	8
Aromatic ligands	6	49
Ligands with heme groups or derivatives	8	15
Ligands with alcohol or acid functional group	13	12

clustering allows the determination of the optimal number of clusters by interactive slicing of the clustering tree. The current test set covers 13 different protein families; therefore, to determine an optimal number of clusters, the following 13 values were tested: 8, 10, 12, 13, 14, 15, 16, 17, 18, 20, 22, 26, and 30. In total, 156 clustering setups (12 clustering schemes with 13 different numbers of clusters) were evaluated, and the Rand statistic scores (see equation (3)) were computed for all cases. The clustering results based on different clustering algorithms and Cavbase scoring functions were very similar overall. Accordingly, virtually all clustering setups were able to distinguish between the 13 different protein families. A cluster analysis based on sequence similarities, with 13 clusters, reveals clear separation of the 13 protein families (data not shown). In sequence space, this result appears self-evident; however, a high level of sequence similarity is not necessarily accompanied by a high level of structural similarity of binding sites. Even entries from the same protein can have dissimilar binding pockets, e.g. if different conformational states are observed.

Two clustering setups (rb/R$_1$ and agglo/R$_2$) achieve an optimal clustering solution, reproducing the external clustering exactly (Table 5). Both Cavbase clusterings show a clear separation in different subfamilies, and all of the 13 clusters contain cavities from only a single enzyme family (Figure 8). If a higher number of clusters is used, enzyme families are separated into different subgroups (e.g. different protein kinase subfamilies, thymidylate kinase), while still separating the different enzyme families. The clustering solutions do not depend strongly on the individual scoring schemes applied, with all three scoring schemes producing consistent results. In the analysis of the current dataset, the R_1 and R_3 scoring schemes perform slightly better than the R_2 scoring scheme. With respect to the various clustering algorithms, the agglo and the partitional methods perform very well, whereas the graph-partitioning (graph) is slightly worse (Table 5). However, 154 of the 156

Table 4. Test set of 105 proteins from 13 diverse protein families used to validate the Cavbase clustering procedure

EC number	SCOP family	SCOP protein	PDB codes
1.1.1.21	Aldo-keto reductases	Aldose reductase	1ads 1ah0 1ah3 1ah4 1az1 1az2 1ef3 1eko 2acr 2acs 2acu
1.1.1.42	Dimeric isocitrate & isopropylmalate dehydrogenases	Isocitrate dehydrogenase	1ai2 1ai3 1bl5 1cw1 1gro 1grp 8icd 9icd
1.14.13.2	FAD-linked reductases, N-terminal domain	*p*-Hydroxybenzoate hydroxylase	1bf3 1cj2 1cj3 1d7l 1ius 1pxa 1pxb
2.7.1.37	Protein kinases, catalytic subunit	Kinases (serine/threonine)	1bkx 1atp 1cdk 1ydr 1hck 1ckp 1b38 1gol 1phk 1csn 1fin 1fin 1erk 3lck 1p38
2.7.1.112	Protein kinases, catalytic subunit	Kinases (tyrosine)	2src 1ir3
2.7.4.9	Nucleotide and nucleoside kinases	Thymidylate kinase	1e9a 1e9b 1e9c 1e9d 1e9e 1gtv 1tmk 2tmk 3tmk 4tmk 5tmp
3.4.21.62	Subtilases	Subtilisin	1au9 1bfu 1bh6 1c3l 1c9j 1sua 1sud 1sue
3.4.23.20	Pepsin-like	Acid protease	1apt 1apu 1apv 1apw 1bxo 1bxq 2wed 3app
4.2.1.1	Carbonic anhydrase	Carbonic anhydrase	1cil 1g52 1g54 1i90 1azm 1bzm 1flj 1keq 1urt
4.4.1.11	Cystathionine synthase-like	Methionine gamma-lyase	1e5e 1e5f
5.3.1.5	Xylose isomerase	D-Xylose isomerase	1xii 1xyc 1xym 5xim 5xin 6xia 6xim 9xia 9xim
5.4.2.1	Cofactor-dependent phosphoglycerate mutase	Phosphoglycerate mutase	1bq3 1bq4 1e58 1e59 1qhf 4pgm
6.3.2.3	Eukaryotic glutathione synthetase	Glutathione synthetases	1glv 1gsa 1gsh 2glt
6.3.2.9	MurCD N-terminal domain	D-Glutamate ligase MurD	1eeh 1uag 2uag 3uag 4uag

The EC code, the SCOP family and protein annotation and the PDB codes for each family are listed.

Table 5. Summary of the 20 best combinations of Cavbase scoring function, cluster algorithm, and number of output cluster according to the Rand statistic score

Rand statistic	Cavbase scoring function	Clustering algorithm	Number of output clusters
1.000	R1	rb	13
1.000	R2	agglo	13
0.998	R3	rb	13
0.997	R3	agglo	14
0.997	R2	agglo	14
0.997	R1	agglo	14
0.997	R1	rb	12
0.997	R1	rbr	13
0.996	R3	agglo	15
0.995	R2	rb	13
0.995	R3	rbr	13
0.994	R3	rb	12
0.994	R2	agglo	12
0.994	R1	agglo	12
0.993	R3	rbr	12
0.993	R1	rbr	12
0.992	R1	agglo	15
0.991	R3	agglo	16
0.991	R1	agglo	13
0.990	R3	agglo	17

clustering setups yield Rand statistic scores greater than 0.90, showing a significant congruence between the Cavbase and the expert classification. On the basis of the results obtained using this test set, and considering the experience with other data sets (see below), it is suggested that the partitional clustering (rb and rbr) or the agglomerative (agglo) algorithms together with either Cavbase scoring function R_1 or R_3 provide optimal results. The latter method tends to produce one very large cluster comprised of cavities from different protein families if the number of output clusters used is too low. However, it performs best in conjunction with a high number of clusters (e.g. greater than 17). As a rule of thumb, the predefined number of presumed clusters should be set close to the expected number of protein families in a dataset.

Obviously, Cavbase is able to differentiate between heterogeneous protein families. In the following case study, the scope of our approach in classifying homologous protein families is investigated using α-carbonic anhydrases and eukaryotic protein kinases.

Classification of α-carbonic anhydrases

At present, the carbonic anhydrase (CA) gene family contains 14 active members. Their basic physiological function is linked to the conversion of carbon dioxide to bicarbonate. They participate in a variety of physiological processes that include pH regulation, carbon dioxide and bicarbonate transport, as well as water and electrolyte balance.[59] CAs originating from the animal kingdom are all of the α-type. The different CA-α isozymes possess a similar architecture of a twisted ten-stranded β sheet. However, they show different levels of sequence identity. The active site is formed by a large cone-shaped cavity with a zinc ion at the bottom. The metal ion is coordinated tetrahedrally by three histidine residues and, most likely due to a pK_a shift, a hydroxide ion. The residues involved in the zinc binding are invariant.[59] Crystal structures are available for six of the 14 families.

According to the SCOP database (version 1.65) a total of 173 CA structures belonging to the CA α superfamily have been deposited with the PDB. The majority of these entries are part of the CA II class. A dataset of 24 catalytic cavities was extracted from the PDB, including examples of all six isozyme classes§. A consistent classification of these entries is obtained *a priori*, by consulting the SCOP and ENZYME databases together with general information about CA isozymes.

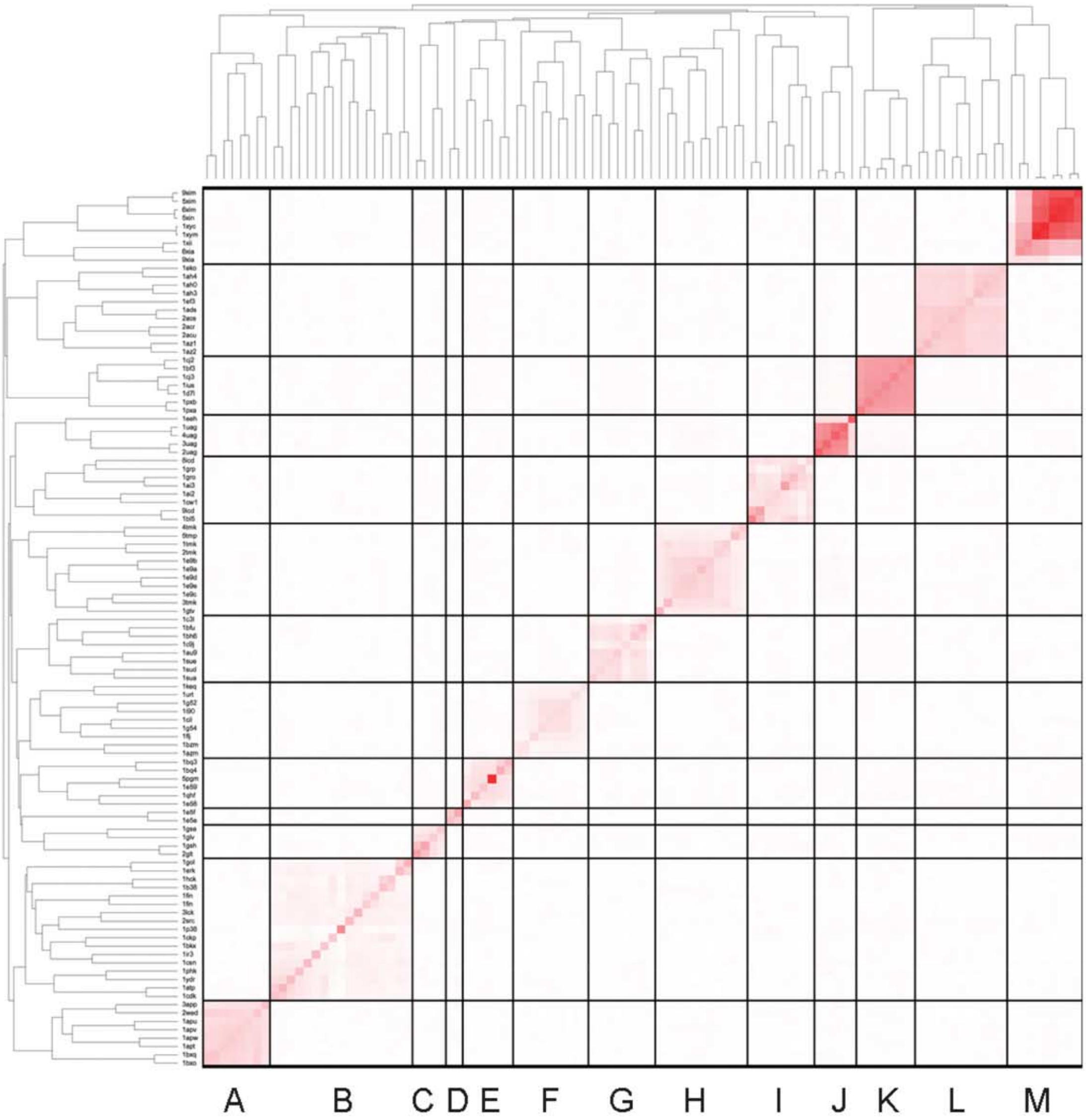
In good agreement with the previously described overall enzyme classification, the results for the CA dataset show the best separation using the R_3 scoring function, together with rb clustering. Nevertheless, the other scoring functions and clustering algorithms also suggest a similar classification (Figure 9). Convincing clustering on the subfamily level is achieved. The classifications obtained differ mainly with respect to the assignment of the only CA-III cavity (PDB code 1flj) in the dataset. By defining a sufficiently large number of expected output clusters, this entry would form a singleton.

Following strategies I, II and III, cavities from different CA-II crystal structures are grouped into two distinct clusters. This separation originates from different conformers adopted by these CA-II entries. Cavities found in the two clusters differ with respect to the conformation of His64. It is known that this residue is mobile and plays an important role in the catalytic mechanism of CAs as a proton shuttle.[60] It can adopt distinct "in" and "out" conformations, and Cavbase distinguishes between the CA-II entries showing the two alternative conformations. In the present dataset, 1bcd and 1ca2 exhibit the out conformation.

Four CA-I cavities are included in the dataset; however, one of them (PDB code 1hcb) performs differently from the other three. It is separate in terms of the Cavbase similarity score from the other three CAs. Seemingly, the automatically extracted cavity for this entry is considerably larger than that for the other three; this is indicated by a very high self-similarity score. Nevertheless, all four combinations of clustering strategies suggest this cavity to be very close to the CA-I cluster formed by 1azm, 1bzm, and 1czm (Figure 9 I–III).

The binding sites of CA-IV (PDB codes 1znc, 2znc, 3znc) exhibit a general similarity to other CA isozymes (e.g. CA-II); however, there are some

§ CA isozymes and corresponding PDB codes used in the present study: CA-I (1azm, 1bzm, 1czm.2, 1hcb), CA-II (1cil, 1g52, 1g54, 1i8z, 1i90, 1a42, 1if4, 1if8, 1bcd, 1ca2), CA-III (1flj), CA-IV (1znc, 2znc, 3znc), CA-V (1dmx, 1dmy, 1urt, 1keq), CA-XII (1jcz, 1jd0).



The clusters comprise the following cavities: cluster A (1apt, 1apu, 1apv, 1apw, 1bxo, 1bxq, 2wed, 3app), cluster B (1bkx, 1atp, 1cdk, 1ydr, 1hck, 1ckp, 1b38, 1gol, 1phk, 1csn, 1fin, 1fin, 1erk, 3lck, 1p38, 2src, 1ir3), cluster C (1glv, 1gsa, 1gsh, 2glt), cluster D (1e5e, 1e5f), cluster E (1bq3, 1bq4, 1e58, 1e59, 1qhf, 5pgm), cluster F (1cil, 1g52, 1g54, 1i90, 1azm, 1bzm, 1flj, 1keq, 1urt), cluster G (1au9, 1bfu, 1bh6, 1c3l, 1c9j, 1sua, 1sud, 1sue), cluster H (1e9a, 1e9b, 1e9c, 1e9d, 1e9e, 1gtv, 1tmk, 2tmk, 3tmk, 4tmk, 5tmp), cluster I (1ai2, 1ai3, 1bl5, 1cw1, 1gro, 1grp, 8icd, 9icd), cluster J (1eeh, 1uag, 2uag, 3uag, 4uag), cluster K (1bf3, 1cj2, 1cj3, 1d7l, 1ius, 1pxa, 1pxb), cluster L (1ads, 1ah0, 1ah3, 1ah4, 1az1, 1az2, 1ef3, 1eko, 2acr, 2acs, 2acu), cluster M (1xii, 1xyc, 1xym, 5xim, 5xin, 6xia, 6xim, 9xia, 9xim)

Figure 8. An optimal Cavbase clustering solution of the enzyme test dataset. Optimal clustering is achieved based on 13 predefined output clusters. The rb clustering algorithm and scoring function R_1 were used. The mutual similarity of the binding cavities, computed by the scoring function R_1, is indicated by the intensity of the red color (dark red, pronounced similarity, white, no similarity). Cavbase separates entries from the different protein families into distinct clusters.

differences. Most notably, the residues of the Val131 to Asp136 loop adopt a conformation pointing towards the solvent in CA-IV, whereas an α-helical conformation is found in other CA isozymes. This helix is directed towards the binding site. A Cavbase comparison of CA-II and CA-IV cavities reveals no similarity in that region. Out of the three CA-IV cavities in the present dataset, one originates from humans (PDB code 1znc) and two are of murine origin (PDB code 2znc and 3znc). They

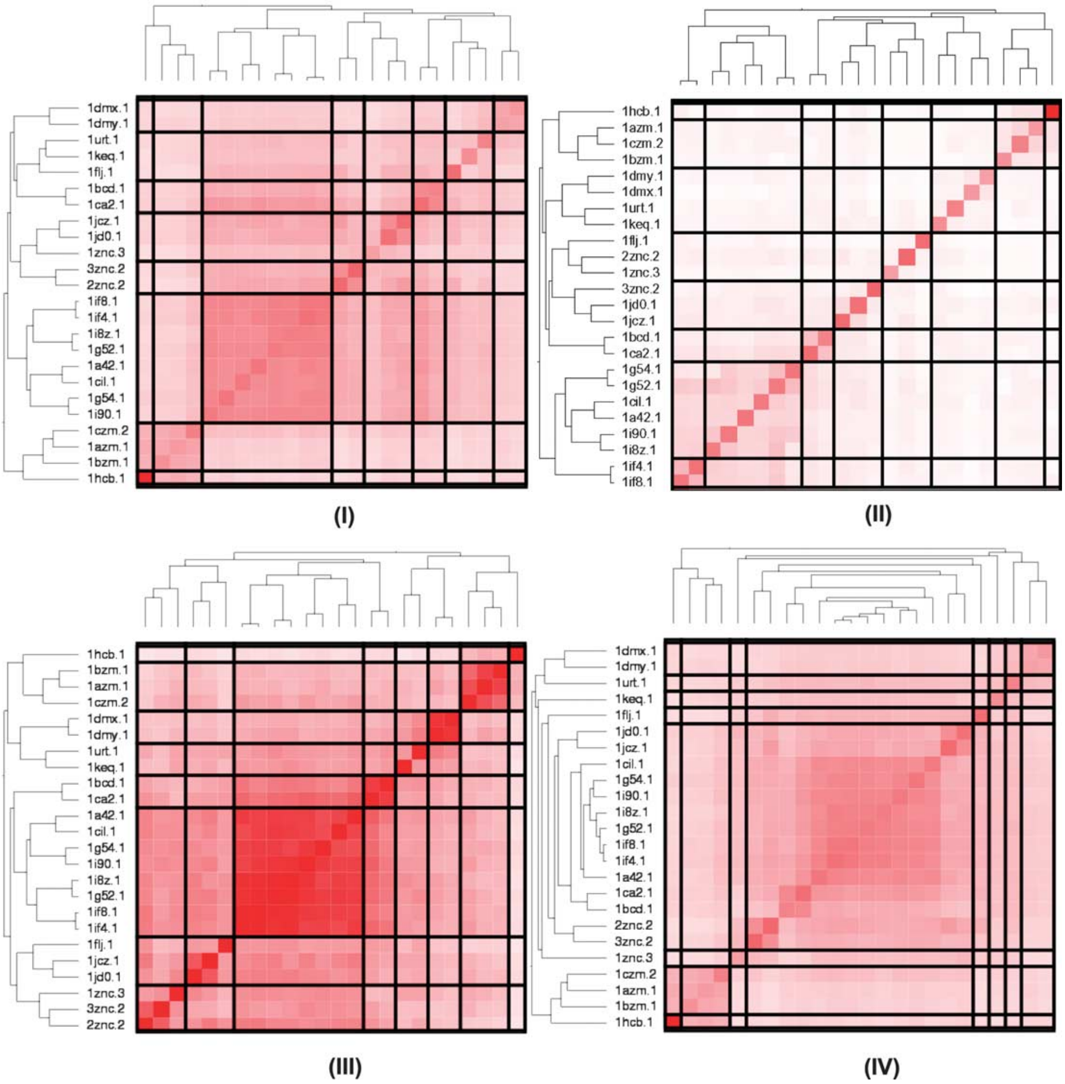


Figure 9. Clustering results for the α-CA isozymes, using four different parameter settings. In all cases, the number of output clusters was set to eight. In (I), (II), and (III), the clustering algorithm rb was used, together with the scoring schemes R_1, R_2, and R_3, respectively. Mutual similarities are expressed by the intensity of the red color (see Figure 8). The different scoring schemes produce consistent results and suggest a reasonable clustering. In (IV), the clustering based on the agglo algorithm, in combination with scoring scheme R_1, is shown. It tends to produce several singletons early and seems to merge many entries into one large cluster. By predefining a larger number of clusters, this large cluster would be decomposed into several smaller clusters.

share a sequence identity of 56%. Several substitutions are found in the active sites (e.g. K91V, M67E, S65T, I141F) of the different species. Furthermore, the cavity from 1znc is significantly smaller than the other two. Despite these deviations, scoring scheme R_3 places all three CA-IV cavities into the same cluster (Figure 9).

Cavbase detects similarities among all four considered CA-V cavities (PDB codes 1dmx, 1dmy, 1urt, 1keq). Depending on the clustering strategy, the CA-V cavities end up in different subclusters (Figure 9). The dataset comprises two wild-type CA-Vs (1dmx and 1dmy) and two double mutants (1keq (F65A/Y131C) and 1urt (Y64H/Y131A)). The areas detected to be similar in the two wild-types and in the wild-type and the mutant cavities are shown in Figure 10. The different physicochemical properties of the mutated amino acids cannot be matched. Nevertheless, all CA-V binding sites have enough similarity to be clustered together.

Classification of protein kinases

Protein kinases form a huge gene family and account for 1.7% of all entries in the human genome.[61] Their catalytic domains share sequence and fold homology (protein kinase fold); nevertheless, they exhibit a rich diversity of regulation modes and substrate specificities.[62–64] The ATP-binding site is located at the interface between the two kinase fold subdomains (lobes). Since

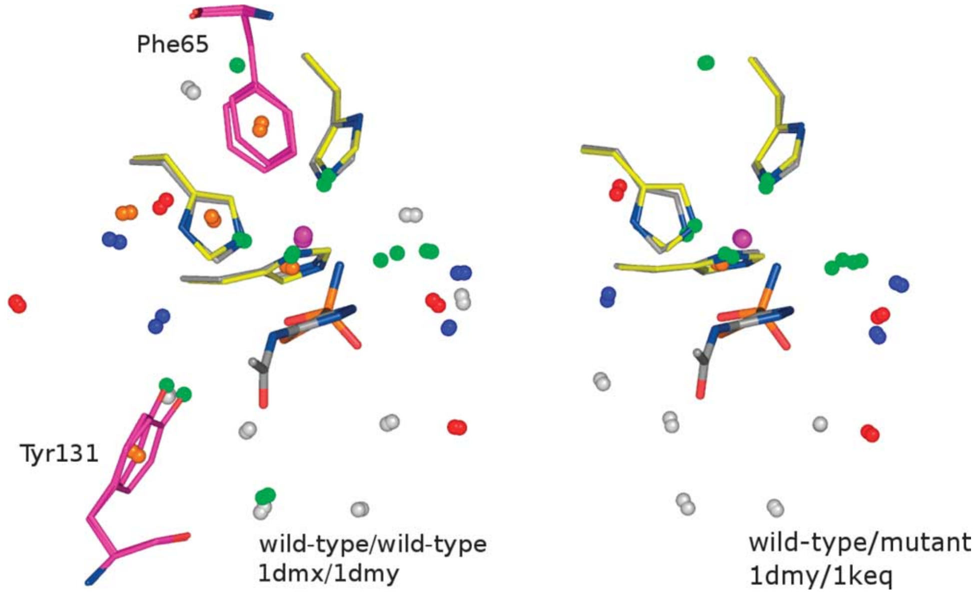


Figure 10. Areas matched between the binding sites of two CA-V wild-type entries (left) (PDB codes 1dmx and 1dmy) and between a wild-type and a mutant isozyme (right) (PDB codes 1dmy and 1keq). For reasons of clarity, only the corresponding pseudocenters, the bound zinc ions (violet spheres) and the sulfonamide inhibitor (1dmy), together with the three histidine residues involved in zinc binding, are shown (color coding as in Figure 2). Additionally, on the left, the phenylalanine and tyrosine residues are displayed (carbon atoms in magenta), that were mutated to alanine and cysteine, respectively. The pseudocenters of the mutated amino acids cannot be matched; but Cavbase still detects pronounced similarities in the binding site.

kinases are involved in the regulation of many cell signaling pathways, they provide attractive targets in disease therapy, e.g. cancer, angiogenesis, neurological diseases, and inflammation.[65–67] In 2001, the first low-molecular-mass ATP-competitive inhibitor (Imatinib, Gleevec©) was introduced to the market as a potent agent against chronic myelogenous leukemia (CML). This success underlines the fact that, even though ATP pockets are rather homologous, they can be addressed selectively by small molecules. Currently, a vast number of medicinal chemistry projects are being carried out in industry to target the ATP-binding site of different kinases with ATP-competitive inhibitors.[67–70] To analyze the performance of our Cavbase cluster analysis, the dataset used by Naumann and Matter[71] was used and extended by additional MAP kinases to comprise 30 kinase cavities in total¶.

The classification obtained by Cavbase, using the R_1 scoring function, the rb clustering method and predefining six clusters, is shown in Figure 11. The clusters (along the diagonal from the bottom-left to the top-right) consist of cavities extracted from the mitogen-activated protein kinases (MAP) of the p38α (cluster A) and Erk2 (cluster B), cyclin-dependent protein kinases (CDKs) and src kinase (cluster C), the fibroblast growth factor receptor kinases and tyrosine kinases (cluster D), the serine/threonine kinase subfamily (cluster E), and the cAMP-dependent protein kinase subfamily (cluster F) (Figure 11). Cavbase is able to separate the different kinase subfamilies automatically, notably considering structural information about the kinase binding sites only, and not about the entire proteins.

Our Cavbase clustering results match well with the landscape analysis of Naumann and Matter using a GRID/cPCA approach. Furthermore, our results are in good agreement with classifications based on CATH or SCOP and with additional information based on multiple sources accomplished by manual intervention. A similar clustering based on sequence similarities separates all 30 entries into different protein kinase subfamilies. However, besides the overall agreement of the clustering in sequence and cavity space, there are some important differences. One notable discrepancy concerns the consideration of distinct activation states of the kinases. These can be captured in cavity space but remain unresolved in sequence space.

For example, the activation of CDKs requires two steps: binding of cognate cyclin, followed by phosphorylation of a threonine residue (Thr160) in the activation loop. Cavbase is able to distinguish between CDKs in the two activation states. The five CDK cavities present in the dataset are grouped into one cluster, which further divides into one

¶ The dataset comprised the kinase ATP binding sites from the following proteins: PKA (1bkx, 1atp, 1cdk, 1bx6, 1stc, 1ydt, 1yds, 1fmo, 1ydr), different Ser/Thr kinases (1phk,1csn), Tyr kinases (1ir3, 1fgi (chain A and chain B), 2src, 3lck), CDK2 (1ckp, 1b38, 1hck, 1fin (chain A and C), and MAP kinases (1gol, 1erk, 1p38, 1pme, 3erk, 4erk, 1bmk, 1a9u,1bl7).

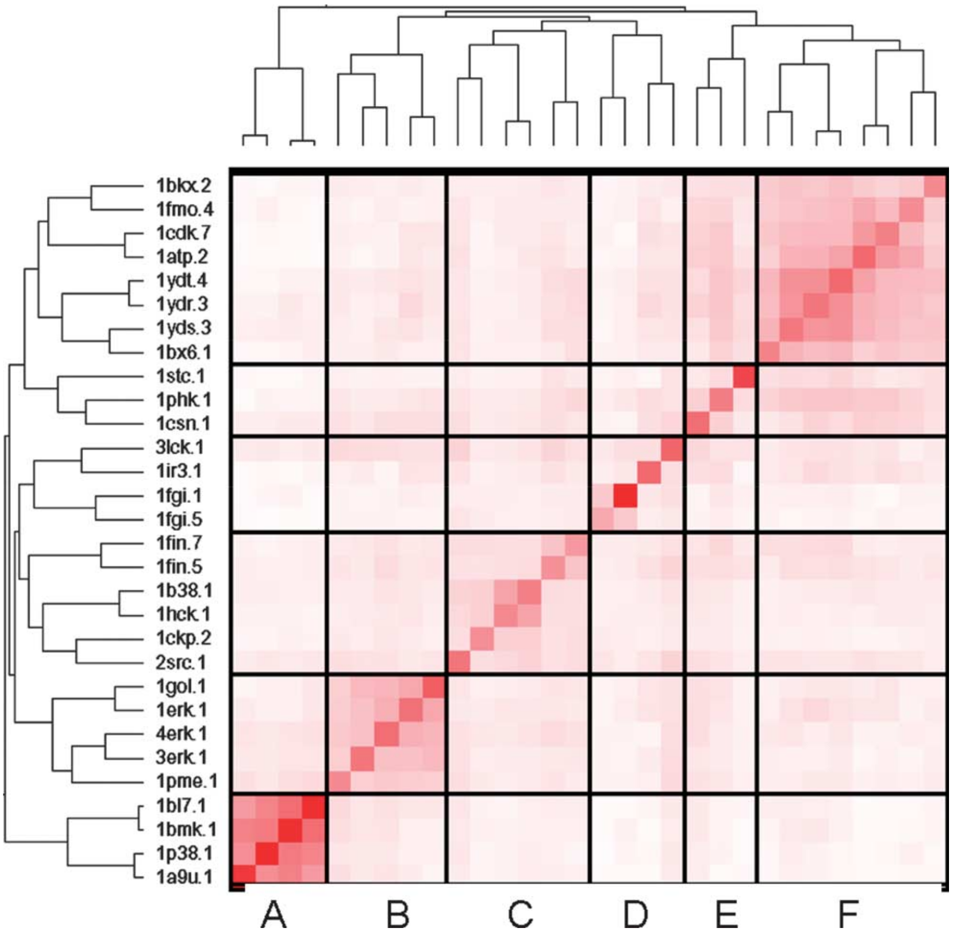


Figure 11. Cavbase clustering for a kinase dataset of 30 cavities (see Figure 8). The R_1 scoring function and the rb clustering algorithm were used to generate six distinct clusters. Cavbase differentiates between the 30 kinases at the subfamily level. The clusters (along the principal diagonal from bottom-left to top-right) comprise cavities from the mitogen-activated protein kinases (MAP) of the (a) p38α and (b) Erk2 subfamilies, (c) the cyclin-dependent protein kinases (CDKs) and src kinase, (d) the fibroblast growth factor receptor kinases and tyrosine kinases, (e) the serine/threonine kinase subfamily, and (f) the cAMP-dependent protein kinase subfamily.

subcluster containing active CDKs (1fin chain A and chain C) and one containing inactive CDKs (1b38, 1ckp, 1hck). Such differences cannot be detected in sequence space.

The Cavbase analysis helps to establish relationships between different protein kinase families in terms of binding site regions that are common and those that differ between families. For example, similarities among the MAP kinase cavities from the Erk2 and p38α subfamilies are detected convincingly. However, differences are also well captured. Figure 12 shows a superposition of two different MAP Erk2 kinase cavities and of a MAP Erk2 and MAP p38-α cavity. The two Erk2 kinases exhibit a high level of similarity. The areas detected as similar comprise almost the entire cavities, including the ATP-binding pockets, and extend towards the activation loop. In contrast, the similarity between the Erk2 and p38-α structures is substantially smaller and limited to the adenine-binding region. In particular, areas next to the hinge-binding region are found to be similar. They are addressed by recurring hydrogen bonding motifs in small-molecule inhibitors targeting the ATP pocket. Accordingly, the Cavbase analysis intuitively helps to locate the selectivity-discriminating regions among binding pockets from related proteins. Of particular interest in this respect are spatial similarities across binding pockets that reside in proteins with low levels of sequence identity. The criteria for comparison in Cavbase focus on the spatial arrangement of physicochemical properties, and not on the actual amino acids along the polypeptide chain. For example, the cavities of two cAMP-dependent kinases (1cdk and 1atp) from pig and mouse are more distinct in sequence space than in cavity space (Figure 11); in the latter they are closely related. The phosphorylase kinase (1phk) shares only a low degree of sequence identity with the cAMP-dependent kinases (~23%). Interestingly, in cavity space its binding pocket shows pronounced similarity to the cAMP-dependent kinases (Figure 13). Furthermore, it shows similarities to the sequentially unrelated casein kinase (1csn).

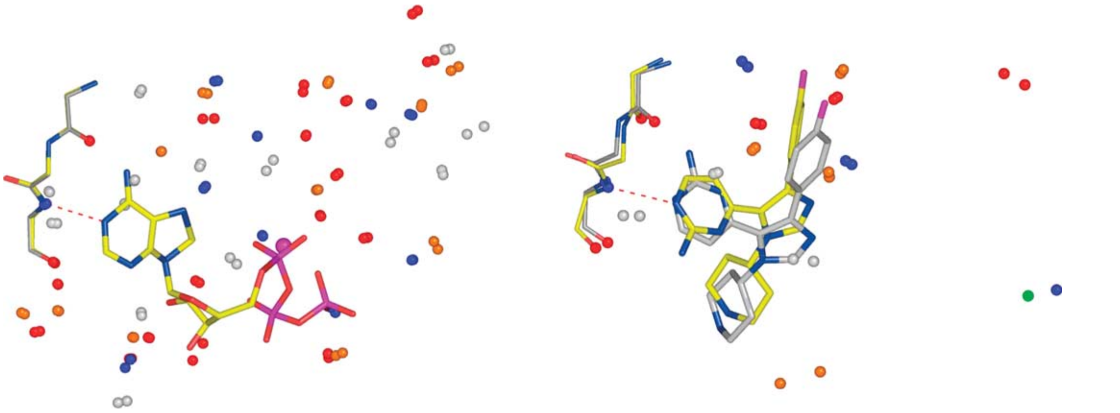


Figure 12. Superposition of MAP kinase binding sites. On the left, a superposition of two Erk2 kinases (PDB code 1erk and 1gol) is shown. Large portions of the binding sites are recognized as being similar. On the right, a superposition of an Erk2 kinase (PDB code 3erk, carbon atoms colored gray) and a MAP p38α kinase (PDB code 1bl7, carbon atoms colored yellow) is displayed. In both pictures, the matching pseudocenters and the hinge backbone protein atoms are displayed (color coding as in Figure 2). Based on the similar hinge binding region, Cavbase superimposes both inhibitors convincingly. In both cases, the hinge coordination *via* the hydrogen bond from the pyrimidine nitrogen atom of the inhibitor to Asp104 (Erk2) and His107 (p38α) is detected.

Such analyses can support the design of selective inhibitors and suggests sets of kinases among which ligand cross-reactivity can be expected.

Conclusion

Here, we present important algorithmic and methodological enhancements of Cavbase. The approach compares and classifies proteins in terms of binding-site exposed physicochemical properties responsible for the recognition of potentially bound ligands. For this purpose, Cavbase assigns to all binding-site residues pseudocenters that encode the recognition properties of the residue's functional groups in the binding pocket. Furthermore, surface patches are assigned to the pseudocenters to measure their accessibility by potentially bound ligands. The quality of the attempted binding pocket comparison in Cavbase relies highly on the completeness, relevance and reliability of the spatial placement of the pseudocenters. To enhance this description, pi pseudocenters representing π interactions putatively performed by the amino acids containing terminal carboxy, carboxamide, and guanidino groups were introduced. Furthermore, the hydrogen bond donor capabilities of cysteine are now considered. In order to account for the edge-to-face interactions of aromatic moieties, the angular parameters for the exposure of pi centers were adjusted. Enhancements to the clique algorithm for the binding site comparisons resulted in a significant reduction of the time needed for binding site analysis, thereby facilitating large-scale comparisons of binding pockets. Additional examples of the discovery of functional similarity of proteins, despite low levels of sequence and fold homology, are presented. In a case study, the potential of Cavbase as a design tool has been demonstrated. Antiviral leads targeting the SARS cysteine protease have been assembled on the basis of the results of a Cavbase similarity search using the individual subpockets of the target protease as input query. Interesting hits are suggested as putative occupants of the protease subpockets. The retrieval of ligands and ligand fragments bound in a similar physicochemical environment permits a better characterization of the SARS protease recognition pockets. Compounds that were selected using this information and subsequently synthesized showed micromolar inhibition of the protease.[84]

Furthermore, it is demonstrated here that Cavbase provides a novel taxonomy to classify proteins. A selected set of cavities is clustered in terms of their mutual cavity similarity. A clustering

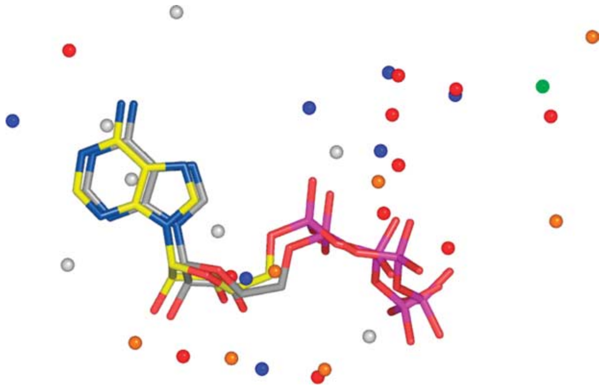


Figure 13. Superposition of phosphorylase kinase (PDB code 1phk) and cAMP-dependent kinase (PDB code 1atp). The matching pseudocenters and bound ATP molecules are displayed (color coding as in Figure 2). The two cavities show extensive similarities in the entire ATP binding pocket, comprising areas next to the hinge region and the adenosine binding site, as well as the DFG motif and parts of the activation loop.

procedure that allows for the fast and automatic classification of large datasets has been implemented. Using a test set of various enzyme structures, optimal settings for clustering could be determined. Cavbase clustering reveals a reliable separation of various dataset entries into individual clusters representing different protein families. The novel technique is applied to two pharmaceutically relevant protein families: the α-carbonic anhydrases and the eukaryotic protein kinases. Cavbase distinguishes successfully between different protein subfamilies and produces a meaningful clustering. In the case of the CAs, Cavbase achieves full separation into single subfamilies. Manually analyzing the indicated relationships across the CA subfamilies elucidates clustering in terms of conformational states and mutational differences. The classification of the protein kinases also resulted in a clear differentiation between the distinct kinase subfamilies. Furthermore, Cavbase succeeds in separating kinase subfamily structures into distinct activation states. This again emphasizes the sensitivity of our approach towards conformational differences among binding pockets. Relationships between kinase subfamilies, such as the Erk2 and p38α subfamilies of the MAP kinases, could be established; regions of the binding sites that are common to the different subfamily cavities, and regions where the essential features differ, are identified. The detection of structurally similar areas of the binding sites of sequentially unrelated kinases helps to identify examples where cross-reactivity due to these structural similarities could be expected.

Materials and Methods

Cavity detection and property description

Cavbase is a method for describing and comparing protein binding pockets in terms of exposed physicochemical properties.[32,72] It is a modular extension of the protein-ligand database Relibase+,[73–75] and the version used in this study contains data for 80,661 binding pockets extracted from 22,885 proteins‖. Protein binding sites are detected using the Ligsite algorithm.[76] The protein under consideration is embedded into a regularly spaced Cartesian grid with 0.5 Å spacing. Any grid points, represented by 1.5 Å diameter probe spheres, that penetrate into the van der Waals sphere of protein atoms are discarded and classified as solvent-inaccessible grid points. For the remaining points, the degree of burial in the protein binding pocket is determined, and neighboring grid points with a high degree of burial are merged together to form contiguous cavities, following the procedure described by Schmitt *et al.*[32] All surface-contacting grid points of such a cluster, apart from the non-buried ones oriented towards the solvent, are used to approximate the cavity surface. Any amino acid with at least one atom closer than 1.1 Å to a surface-contacting grid point is defined as a cavity flanking residue.

The physicochemical properties of the amino acid residues flanking the cavity are encoded in terms of pseudocenters, which represent appropriate 3D descriptors. Each of these pseudocenters is defined as a point in 3D space and is associated with one of the following properties determinant for molecular recognition: hydrogen bond (HB) donor, HB acceptor, mixed HB donor/acceptor, hydrophobic contact, aliphatic contact or aromatic contact. This condensed representation allows for efficient similarity searching on the basis of a reduced set of input variables.

In a subsequent filtering step, only the pseudocenters that can expose their property onto the cavity surface are retained. Two vectors, **v** and **r**, are calculated for each pseudocenter. The first vector, **v**, reflects the mean orientation along which a particular interaction could be formed, whereas the second vector, **r**, points towards the cavity surface. The angle enclosed by the two vectors serves as a criterion for determining whether a particular pseudocenter is considered in the analysis or discarded. Pseudocenters for which this angle is greater than the predefined cut-off value (donor and acceptor, 100°; donor/acceptor, 120°; aromatic (pi), 60°)[32] are considered unable to expose their property towards the cavity surface, and are discarded from the set of pseudocenters that define the cavity. Finally, all surface-contacting grid points describing the cavity surface are assigned to the nearest pseudocenter (providing this is within 3 Å). Thus, the properties of the pseudocenters are mapped onto surface patches describing the exterior of the cavity.

Increasing the speed of computational binding site comparisons

One prerequisite for the large-scale clustering analysis of protein cavities is an algorithm that is able to compute similarities between binding sites in a fast and efficient way. In principle, Cavbase can perform such comparisons. However, even though a mutual comparison of two binding cavities is easy to compute, a large-scale comparison across huge cavity datasets will be very demanding and hardly feasible. A clique algorithm is applied to detect common substructures between two cavities.[77] The clique algorithm operates on the product graph P, which has to be constructed beforehand. A node (u,v), consisting of a pair of pseudocenters from two binding sites, is inserted into P if the two pseudocenters have comparable labels (pseudocenter types).[32] Two nodes (u,v) and (u',v') in P are connected if the corresponding pseudocenter pairs fulfill the following conditions:

$$\text{(i)} \qquad d(u,v) \le d_{\max} \wedge d(u',v') \le d_{\max}$$

the distance between both pseudocenters in each cavity has to be smaller than $d_{\max}$ (default value: 12 Å, thus considering, in particular, local patterns) and:

$$\text{(ii)} \qquad |d(u,v) - d(u',v')| \le d_{\text{diff}}$$

the difference between both distances has to be smaller than a predefined tolerance value d_{diff} (2 Å in the original approach). Since the computing time of the clique algorithm depends strongly on the number of connected nodes in the product graph P, one strategy to accelerate binding site comparisons is the reduction of the number

‖ Relibase is accessible on the web from http://relibase.ccdc.cam.ac.uk, http://relibase.ebi.ac.uk, or http://relibase.rutgers.edu. Relibase+ and Cavbase are distributed by the Cambridge Crystallographic Data Centre (CCDC), Cambridge, UK, www.ccdc.cam.ac.uk

Table 6. During the clique detection the shorter distance of two pseudocenter pairs is used to adjust the value for the acceptance tolerance (d_{diff}) in a distance-dependent manner

Shorter distance between the two pseudocenters pairs (Å)	Tolerance distance d_{diff} (Å)
10.0–12.0	2.0
8.0–10.0	1.6
4.0–8.0	1.2
2.0–4.0	0.8
<2.0	0.6

of connected nodes in P. The tolerance value d_{diff} is therefore assigned in a distance-dependent fashion. The shorter distance of the two pseudocenter pairs is used to adjust the value for d_{diff}. It adopts lower values if at least one of the distances is small, thereby reducing the number of connected nodes.

A predefined number (default 100) of the largest clique solutions is accepted and further evaluated by scoring them according to the degree of spatial overlap between corresponding cavity surface patches. The following scoring procedure is applied to every accepted clique solution: The two binding sites are superimposed based on the matching pseudocenters found in the clique detection (Table 6). In addition to the matching pseudocenters detected in the clique detection, every pseudocenter pair assigned to the same physicochemical property is analyzed in terms of the surface patch overlap. This is done to potentially improve the initial clique solutions. The surface patch overlap for each pseudocenter pair is calculated by summing the relative frequency of surface grid points of both surface patches that fall next to each other below a distance threshold of 1.0 Å. A match is considered to contribute to the overall binding site similarity only if at least 70% of the surface points of the matching surface patches are shared in common, to avoid the consideration of strongly fragmented surface patches. The R_1 scoring value for the comparison of two binding sites is calculated by summing all surface patch overlap values of the pseudocenter pairs that pass the overlap criterion. These pseudocenters represent the physicochemical properties shared by both cavities. Only the most highly scored clique solution is considered further.[32] Approximately 80% of the computational effort in a binding site comparison is spent on the scoring process. Accordingly, we sought to accelerate the evaluation of the degree of overlap of two surface patches. Since the surface patch overlap is determined for all pseudocenter pairs assigned to the same type, it is even determined if both patches are very distant from each other and cannot be superimposed fulfilling the criteria mentioned above. Obviously, this is the case if the distance between the corresponding pseudocenters is too large; these comparisons have to be avoided. Therefore, only the surface patches corresponding to pseudocenters that fall within 4 Å of each other are considered further. Using these heuristics, the speed of a binding site comparison could be enhanced 40-fold.

Two alternative scoring schemes have been used to rank the best scored superposition according to R_1. Scoring scheme R_2 (equation (1)) reflects, in addition, the root-mean-square deviations (RMSD) of the coordinates of the n matching pseudocenters.[32] It disfavors fragmented non-contiguous clique solutions that obtain an artificially high R_1 score. Scoring scheme R_3 (equation (2)) is calculated analogously to the Tanimoto index.[78] It accounts for the number of matched pseudocenters, n_{match}, but normalizes the score with respect to the total size of the cavities, expressed by the total number of pseudocenters n_{pseu1} and n_{pseu2}, respectively.

$$R_2 = \frac{R_1 - 0.7n}{\text{RMSD}} \tag{1}$$

$$R_3 = \frac{n_{\text{match}}}{n_{\text{pseu1}} + n_{\text{pseu2}} - n_{\text{match}}} \tag{2}$$

This latter score is further normalized to a value in the range zero to 1; accordingly, it adopts a value of 1 if a cavity is compared with itself.

For the graphical analysis of cavities and superpositions of two cavities, Cavbase has been equipped with an interface to Pymol[a].

Cavity clustering procedure

Each cavity in the dataset is compared to all other cavities. For each one-to-one comparison, Cavbase returns three similarity scores, according to the three schemes described above. The resulting scores are stored in a similarity matrix that serves as input for various clustering algorithms. Four different clustering algorithms, as implemented in the clustering toolkit CLUTO[b], were used. These consisted of two partitional (rb and rbr), one agglomerative (agglo), and one graph-partitioning (graph) method. The rb and rbr methods split the dataset into two groups. These groups are then repeatedly split further, until a predefined number of output clusters is obtained. The rbr method incorporates, in addition to the procedure of the rb method, a final global optimization step. The agglo method initially regards each single object as an individual cluster ("singleton"), and sequentially merges the two most similar clusters into a new cluster. The similarity of the newly created cluster with respect to all the others is calculated using the unweighted pair group method with arithmetic mean (UPGMA) approach, which evaluates the distance between two clusters by averaging over the similarities of all the pairs of objects that can be formed between the two clusters. In the graph method, the objects are described as nodes of a graph. Nodes are connected if they are nearest neighbors. During clustering the graph is dissected into subgraphs, whilst attempting to minimize the number of edges that are broken. In combination with the different clustering algorithms, agglomerative merging schemes can be applied to obtain a hierarchical tree structure for the different clustering solutions. This allows for an intuitive navigation through the clustered solutions and supports the detection of relationships. All methods were used with the default settings provided by CLUTO.

In total, 12 different combinations of clustering options were evaluated, combining four different clustering algorithms with three different Cavbase scoring schemes. In this initial validation study, the quality of the different Cavbase clusterings has to be assessed. To allow for an independent assessment of the quality and relevance of the obtained clustering of our method, we considered enzyme structures: In this case, we can refer to the

[a] http://pymol.sourceforge.net/

[b] G. Karypis (2002). CLUTO—a clustering toolkit (release 2.1 http://glaros.dtc.umn.edu/gkhome/views/cluto). Technical Report 02-017, University of Minnesota, Department of Computer Science.

enzyme classification (EC) code to estimate the function of a protein.[79] The four-part EC code contains information about the catalyzed reaction and the substrates or cofactors used, and therefore classifies enzymes according to their biochemical function. To compare the quality of a cluster model A with a reference model B, so-called external evaluation measures can be used.[80,81] In our case, the reference model B corresponds to a grouping defined by the external (expert) classification in terms of EC numbers.

A well-known and frequently used evaluation measure is the Rand statistic, which is defined as follows:

$$Q(A,B) = \frac{N_{\text{ss}} + N_{\text{dd}}}{N_{\text{ss}} + N_{\text{sd}} + N_{\text{ds}} + N_{\text{dd}}} \tag{3}$$

where N_{ss} is the number of pairs of elements (x,y) that are put in the same cluster in both A and B. Likewise, N_{sd} is the number of pairs (x,y) that are put in the same cluster in A but in different clusters in B, N_{ds} is the number of pairs (x,y) that are put in different clusters in A but in the same cluster in B, and N_{dd} is the number of pairs (x,y) that are put in different clusters in both A and B. In other words, $Q(A,B)$ is simply the fraction of pairs on which the two partitions A and B agree. The Rand statistic scores can adopt values between zero and 1, where higher values mean a higher level of similarity between A and B.

SCOP-based and sequence-based clustering

The Cavbase classification was compared to sequence-based and SCOP(Structural Classification of Proteins, version 1.65)-based classification schemes, in order to assess the relevance of our clustering. The SCOP database classifies proteins using sequence-comparison and fold-comparison techniques; however, this is accomplished by manual annotation.[82] A hierarchical classification is constructed to reveal relationships between different proteins. Since SCOP classifies proteins at the domain level, the corresponding domain entries for the protein cavities were determined, and the SCOP superfamily and family annotations were used to assess the Cavbase classification.

For the sequence-based clustering, the sequences of the considered proteins were extracted using Relibase+. Subsequently, they were mutually aligned using FASTA 3.5 with standard settings.[83] Sequence identity values provided by FASTA were normalized to values between zero and 1, and used to construct a similarity matrix. Subsequently, the same clustering procedures as those used for the Cavbase similarity scores were applied.

Cavity datasets used for similarity searching

Cavbase detects multiple depressions on the protein surface that might serve as ligand binding sites. To focus on relevant binding sites only, the following pragmatic filter was applied. Only cavities with a bound ligand comprising between five and 75 non-hydrogen atoms were considered further. This filter enhances the probability that the sites considered are relevant to the binding of small-molecule ligands. Such sites consist, in particular, of catalytic cavities, which usually have co-factors, substrates or inhibitors bound. Applying these criteria to the Spring 2003 version of the PDB database returned 9446 binding pockets (dataset I).

Drugscore analysis and contour level adjustment

To validate the spatial assignment of physicochemical properties in Cavbase, a validation set of 214 protein complexes was compiled. The bound ligand was extracted from the original PDB protein file, stored in MOL2 format and used as input for Drugscore (version 1.2). The binding site was defined as being comprised by the residues surrounding the bound ligand within a 6 Å radius. To allow for a comparison of the binding site properties in terms of the exposed physicochemical properties described by Cavbase or Drugscore, the original grid defined by Ligsite was used to calculate Drugscore hotspots.[40] Since the same grid has been used for both evaluations, a superposition of the results of both binding analyses is given. For each protein in the validation set, Drugscore hotspots were calculated and compared visually with the Cavbase surface patch description. Drugscore hotspots were scaled to values between zero and 100, where 100 denotes areas of favorable interaction. To allow for a comparison of the different hotspot fields, an iterative procedure was applied to determine an appropriate contour level for each probe. Starting from a contour level of 100, this value was decreased gradually until 0.6% of the total grid points were considered for contouring. This level was adjusted empirically and revealed contoured regions that are probably larger than those used for a hotspot analysis in inhibitor design. However, the hotspot analysis indicates regions in space suited to accommodating a ligand functional group.

Acknowledgements

The authors thank Simon Cottrell for carefully reading the manuscript.

References

1. Martin, A. C., Orengo, C. A., Hutchinson, E. G., Jones, S., Karmirantzou, M., Laskowski, R. A. *et al*. (1998). Protein folds and functions. *Structure*, **6**, 875–884.
2. Orengo, C. A., Sillitoe, I., Reeves, G. & Pearl, F. M. (2001). Review: what can structural classifications reveal about protein evolution? *J. Struct. Biol.* **134**, 145–165.
3. Nagano, N., Orengo, C. A. & Thornton, J. M. (2002). One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J. Mol. Biol.* **321**, 741–765.
4. Anantharaman, V., Aravind, L. & Koonin, E. V. (2003). Emergence of diverse biochemical activities in evolutionarily conserved structural scaffolds of proteins. *Curr. Opin. Chem. Biol.* **7**, 12–20.
5. Weber, A., Casini, A., Heine, A., Kuhn, D., Supuran, C. T., Scozzafava, A. & Klebe, G. (2004). Unexpected nanomolar inhibition of carbonic anhydrase by COX-2-selective celecoxib: new pharmacological opportunities due to related binding site recognition. *J. Med. Chem.* **47**, 550–557.
6. 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.

7. Spriggs, R. V., Artymiuk, P. J. & Willett, P. (2003). Searching for patterns of amino acids in 3D protein structures. *J. Chem. Inf. Comput. Sci.* **43**, 412–421.
8. 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.
9. Kleywegt, G. J. (1999). Recognition of spatial motifs in protein structures. *J. Mol. Biol.* **285**, 1887–1897.
10. Hamelryck, T. (2003). Efficient identification of side-chain patterns using a multidimensional index tree. *Proteins: Struct. Funct. Genet.* **51**, 96–108.
11. Bachar, O., Fischer, D., Nussinov, R. & Wolfson, H. (1993). A computer vision based technique for 3-D sequence-independent structural comparison of proteins. *Protein Eng.* **6**, 279–288.
12. Fischer, D., Wolfson, H. & Nussinov, R. (1993). Spatial, sequence-order-independent structural comparison of alpha/beta proteins: evolutionary implications. *J. Biomol. Struct. Dynam.* **11**, 367–380.
13. Fischer, D., Norel, R., Wolfson, H. & Nussinov, R. (1993). Surface motifs by a computer vision technique: searches, detection, and implications for protein–ligand recognition. *Proteins: Struct. Funct. Genet.* **16**, 278–292.
14. 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.
15. Fischer, D., Lin, S. L., Wolfson, H. L. & Nussinov, R. (1995). A geometry-based suite of molecular docking processes. *J. Mol. Biol.* **248**, 459–477.
16. 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.
17. 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.
18. Lin, S. L. & Nussinov, R. (1996). Molecular recognition *via* face center representation of a molecular surface. *J. Mol. Graph.* **78–90**, 95–97.
19. Norel, R., Lin, S. L., Wolfson, H. J. & Nussinov, R. (1994). Shape complementarity at protein–protein interfaces. *Biopolymers*, **34**, 933–940.
20. Rosen, M., Lin, S. L., Wolfson, H. & Nussinov, R. (1998). Molecular shape comparisons in searches for active sites and functional similarity. *Protein Eng.* **11**, 263–277.
21. Shulman-Peleg, A., Nussinov, R. & Wolfson, H. J. (2004). Recognition of functional sites in protein structures. *J. Mol. Biol.* **339**, 607–633.
22. Pennec, X. & Ayache, N. (1998). A geometric algorithm to find small but highly similar 3D substructures in proteins. *Bioinformatics*, **14**, 516–522.
23. Russell, R. B. (1998). Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J. Mol. Biol.* **279**, 1211–1227.
24. Stark, A. & Russell, R. B. (2003). Annotation in three dimensions. PINTS: patterns in non-homologous tertiary structures. *Nucl. Acids Res.* **31**, 3341–3344.
25. Kinoshita, K., Furui, J. & Nakamura, H. (2002). Identification of protein functions from a molecular surface database, eF-site. *J. Struct. Funct. Genomics*, **2**, 9–22.
26. Kinoshita, K. & Nakamura, H. (2003). Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. *Protein Sci.* **12**, 1589–1595.
27. Poirrette, A. R., Artymiuk, P. J., Rice, D. W. & Willett, P. (1997). Comparison of protein surfaces using a genetic algorithm. *J. Comput. Aided Mol. Des.* **11**, 557–569.
28. Lehtonen, J. V., Denessiouk, K., May, A. C. & 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. Pickering, S. J., Bulpitt, A. J., Efford, N., Gold, N. D. & Westhead, D. R. (2001). AI-based algorithms for protein surface comparisons. *Comput. Chem.* **26**, 79–84.
30. Wangikar, P. P., Tendulkar, A. V., Ramya, S., Mali, D. N. & Sarawagi, S. (2003). Functional sites in protein families uncovered *via* an objective and automated graph theoretic approach. *J. Mol. Biol.* **326**, 955–978.
31. Jambon, M., Imberty, A., Deleage, G. & Geourjon, C. (2003). A new bioinformatic approach to detect common 3D sites in protein structures. *Proteins: Struct. Funct. Genet.* **52**, 137–145.
32. Schmitt, S., Kuhn, D. & Klebe, G. (2002). A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.* **323**, 387–406.
33. 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.
34. Allen, F. H. (2002). The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallog. sect. B*, **58**, 380–388.
35. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
36. McDonald, I. K. & Thornton, J. M. (1994). Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **238**, 777–793.
37. Bartlett, G. J., Porter, C. T., Borkakoti, N. & Thornton, J. M. (2002). Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.* **324**, 105–121.
38. Mitchell, J. B., Nandi, C. L., McDonald, I. K., Thornton, J. M. & Price, S. L. (1994). Amino/aromatic interactions in proteins: is the evidence stacked against hydrogen bonding? *J. Mol. Biol.* **239**, 315–331.
39. Meyer, E. A., Castellano, R. K. & Diederich, F. (2003). Interactions with aromatic rings in chemical and biological recognition. *Angew. Chem. Int. Ed Engl.* **42**, 1210–1250.
40. Gohlke, H., Hendlich, M. & Klebe, G. (2000). Predicting binding modes, binding affinities and "hot spots" for protein–ligand complexes using a knowledge-based scoring function. *Persp. Drug Des. Discov.* **20**, 115–144.
41. Gohlke, H. & Klebe, G. (2001). Statistical potentials and scoring functions applied to protein–ligand binding. *Curr. Opin. Struct. Biol.* **11**, 231–235.
42. Gohlke, H., Hendlich, M. & Klebe, G. (2000). Knowledge-based scoring function to predict protein–ligand interactions. *J. Mol. Biol.* **295**, 337–356.

43. Nissink, J. W., Murray, C., Hartshorn, M., Verdonk, M. L., Cole, J. C. & Taylor, R. (2002). A new test set for validating predictions of protein–ligand interaction. *Proteins: Struct. Funct. Genet.* **49**, 457–471.
44. Russell, R. B., Sasieni, P. D. & Sternberg, M. J. (1998). Supersites within superfolds. Binding site similarity in the absence of homology. *J. Mol. Biol.* **282**, 903–918.
45. Branden, C. & Tooze, J. (1999). An example of enzyme catalysis: serine proteases. *Introduction to Protein Structure* 2nd edit., pp. 205–220, Garland, New York (chapt. 11).
46. Wallace, A. C., Laskowski, R. A. & Thornton, J. M. (1996). Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. *Protein Sci.* **5**, 1001–1013.
47. Thoden, J. B., Frey, P. A. & Holden, H. M. (1996). Molecular structure of the NADH/UDP-glucose abortive complex of UDP-galactose 4-epimerase from *Escherichia coli*: implications for the catalytic mechanism. *Biochemistry,* **35**, 5137–5144.
48. Lupas, A. N., Ponting, C. P. & Russell, R. B. (2001). On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *J. Struct. Biol.* **134**, 191–203.
49. Fouchier, R. A., Kuiken, T., Schutten, M., Amerongen, G. v., Doornum, G. J. v., Hoogen, B. G. v. d. *et al.* (2003). Aetiology: Koch's postulates fulfilled for SARS virus. *Nature*, **423**, 240.
50. Kuiken, T., Fouchier, R. A., Schutten, M., Rimmelzwaan, G. F., Amerongen, G. v., Riel, D. v. *et al.* (2003). Newly discovered coronavirus as the primary cause of severe acute respiratory syndrome. *Lancet*, **362**, 263–270.
51. Ziebuhr, J., Snijder, E. J. & Gorbalenya, A. E. (2000). Virus-encoded proteinases and proteolytic processing in the Nidovirales. *J. Gen. Virol.* **81**, 853–879.
52. Anand, K., Palm, G. J., Mesters, J. R., Siddell, S. G., Ziebuhr, J. & Hilgenfeld, R. (2002). Structure of coronavirus main proteinase reveals combination of a chymotrypsin fold with an extra alpha-helical domain. *EMBO J.* **21**, 3213–3224.
53. Anand, K., Ziebuhr, J., Wadhwani, P., Mesters, J. R. & Hilgenfeld, R. (2003). Coronavirus main proteinase (3CLpro) structure: basis for design of anti-SARS drugs. *Science*, **300**, 1763–1767.
54. Yang, H., Yang, M., Ding, Y., Liu, Y., Lou, Z., Zhou, Z. *et al.* (2003). The crystal structures of severe acute respiratory syndrome virus main protease and its complex with an inhibitor. *Proc. Natl Acad. Sci. USA*, **100**, 13190–13195.
55. Matthews, D. A., Dragovich, P. S., Webber, S. E., Fuhrman, S. A., Patick, A. K., Zalman, L. S. *et al.* (1999). Structure-assisted design of mechanism-based irreversible inhibitors of human rhinovirus 3C protease with potent antiviral activity against multiple rhinovirus serotypes. *Proc. Natl Acad. Sci. USA*, **96**, 11000–11007.
56. Dragovich, P. S., Prins, T. J., Zhou, R., Webber, S. E., Marakovits, J. T., Fuhrman, S. A. *et al.* (1999). Structure-based design, synthesis, and biological evaluation of irreversible human rhinovirus 3C protease inhibitors. 4. Incorporation of P1 lactam moieties as L-glutamine replacements. *J. Med. Chem.* **42**, 1213–1224.
57. Shie, J. J., Fang, J. M., Kuo, T. H., Kuo, C. J., Liang, P. H., Huang, H. J. *et al.* (2005). Inhibition of the severe acute respiratory syndrome 3CL protease by peptidomimetic alpha,beta-unsaturated esters. *Bioorg. Med. Chem.* **13**, 5240–5252.
58. Yang, H., Xie, W., Xue, X., Yang, K., Ma, J., Liang, W. *et al.* (2005). Design of wide-spectrum inhibitors targeting coronavirus main proteases. *PLoS. Biol.* **3**, e324.
59. Lindskog, S. (1997). Structure and mechanism of carbonic anhydrase. *Pharmacol. Ther.* **74**, 1–20.
60. Kim, C. Y., Whittington, D. A., Chang, J. S., Liao, J., May, J. A. & Christianson, D. W. (2002). Structural aspects of isozyme selectivity in the binding of inhibitors to carbonic anhydrases II and IV. *J. Med. Chem.* **45**, 888–893.
61. Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. (2002). The protein kinase complement of the human genome. *Science*, **298**, 1912–1934.
62. Nolen, B., Taylor, S. & Ghosh, G. (2004). Regulation of protein kinases; controlling activity through activation segment conformation. *Mol. Cell*, **15**, 661–675.
63. Engh, R. A. & Bossemeyer, D. (2002). Structural aspects of protein kinase control-role of conformational flexibility. *Pharmacol. Ther.* **93**, 99–111.
64. Huse, M. & Kuriyan, J. (2002). The conformational plasticity of protein kinases. *Cell*, **109**, 275–282.
65. Saklatvala, J. (2004). The p38 MAP kinase pathway as a therapeutic target in inflammatory disease. *Curr. Opin. Pharmacol.* **4**, 372–377.
66. Matter, A. (2001). Tumor angiogenesis as a therapeutic target. *Drug Discov. Today*, **6**, 1005–1024.
67. Dancey, J. & Sausville, E. A. (2003). Issues and progress with protein kinase inhibitors for cancer treatment. *Nature Rev. Drug Discov.* **2**, 296–313.
68. Cohen, P. (2002). Protein kinases—the major drug targets of the twenty-first century? *Nature Rev. Drug Discov.* **1**, 309–315.
69. Traxler, P. M. (1998). Tyrosine kinase inhibitors in cancer treatment (part II). *Exp. Opin. Ther. Patents*, **8**, 1599–1625.
70. Noble, M. E., Endicott, J. A. & Johnson, L. N. (2004). Protein kinase inhibitors: insights into drug design from structure. *Science*, **303**, 1800–1805.
71. Naumann, T. & Matter, H. (2002). Structural classification of protein kinases using 3D molecular interaction field analysis of their ligand binding sites: target family landscapes. *J. Med. Chem.* **45**, 2366–2378.
72. Schmitt, S., Hendlich, M. & Klebe, G. (2001). From structure to function: a new approach to detect functional similarity among proteins independent from sequence and fold homology. *Angew. Chem. Int. Ed. Engl.* **40**, 3141–3144.
73. Hendlich, M. (1998). Databases for protein–ligand complexes. *Acta Crystallog. sect. D*, **54**, 1178–1182.
74. Hendlich, M., Bergner, A., Günther, J. & Klebe, G. (2003). Relibase: design and development of a database for comprehensive analysis of protein–ligand interactions. *J. Mol. Biol.* **326**, 607–620.
75. Gunther, J., Bergner, A., Hendlich, M. & Klebe, G. (2003). Utilising structural knowledge in drug design strategies: applications using Relibase. *J. Mol. Biol.* **326**, 621–636.
76. Hendlich, M., Rippmann, F. & Barnickel, G. (1997). LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph. Model*, **359–63**, 389.
77. Bron, C. & Kerbosch, J. (1973). Algorithm 457. Finding all cliques of an undirected graph. *Commun. ACM*, **16**, 575–577.

78. Godden, J. W., Xue, L., Stahura, F. L. & Bajorath, J. (2000). Searching for molecules with similar biological activity: analysis by fingerprint profiling. *Pac. Symp. Biocomput.*, 566–575.
79. Bairoch, A. (2000). The ENZYME database in 2000. *Nucl. Acids Res*, **28**, 304–305.
80. Halkidi, M., Batistakis, Y. & Vazirgiannis, M. (2002). Cluster validity methods: part I. *SIGMOD Record*, **31(2)**, 40–45.
81. Halkidi, M., Batistakis, Y. & Vazirgiannis, M. (2002). Cluster validity methods: part II. *SIGMOD Record*, **31(3)**, 19–27.
82. Lo Conte, L., Brenner, S. E., Hubbard, T. J., Chothia, C. & Murzin, A. G. (2002). SCOP database in 2002: refinements accommodate structural genomics. *Nucl. Acids Res*, **30**, 257–264.
83. Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
84. Al-Gharabli, S.I., Shah, S.T.A., Weik, S., Schmidt, B.M.F., Mesters, J.R., Kuhn, D. *et al.* (2006). An efficient method for the synthesis of peptide aldehyde libraries employed in the discovery of reversible SARS corona virus main protease (SARS–CoV Mpro) inhibitors. *Chem Bio Chem*, in press.

Edited by Michael J. Sternberg

(Received 17 February 2006; received in revised form 31 March 2006; accepted 6 April 2006)
Available online 25 April 2006