PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Cofactor-binding sites in proteins of deviating sequence: Comparative analysis and clustering in torsion angle, cavity, and fold space

Björn Stegemann[†] and Gerhard Klebe[*]

Institut für Pharmazeutische Chemie, Philipps-Universität Marburg, Marbacher Weg 6, D-35032 Marburg, Germany

ABSTRACT

Small molecules are recognized in protein-binding pockets through surface-exposed physicochemical properties. To optimize binding, they have to adopt a conformation corresponding to a local energy minimum within the formed protein–ligand complex. However, their conformational flexibility makes them competent to bind not only to homologous proteins of the same family but also to proteins of remote similarity with respect to the shape of the binding pockets and folding pattern. Considering drug action, such observations can give rise to unexpected and undesired cross reactivity. In this study, datasets of six different cofactors (ADP, ATP, NAD(P)(H), FAD, and acetyl CoA, sharing an adenosine diphosphate moiety as common substructure), observed in multiple crystal structures of protein–cofactor complexes exhibiting sequence identity below 25%, have been analyzed for the conformational properties of the bound ligands, the distribution of physicochemical properties in the accommodating protein-binding pockets, and the local folding patterns next to the cofactor-binding site. State-of-the-art clustering techniques have been applied to group the different protein–cofactor complexes in the different spaces. Interestingly, clustering in cavity (Cavbase) and fold space (DALI) reveals virtually the same data structuring. Remarkable relationships can be found among the different spaces. They provide information on how conformations are conserved across the host proteins and which distinct local cavity and fold motifs recognize the different portions of the cofactors. In those cases, where different cofactors are found to be accommodated in a similar fashion to the same fold motifs, only a commonly shared substructure of the cofactors is used for the recognition process.

Proteins 2011; 00:000–000.

Key words: protein–ligand complexes; clustering of proteins; analysis of torsion angle space; analysis of binding cavity space; analysis of local folding patterns; cofactor-binding sites; drug cross reactivity; statistical analysis of crystal structure data.

INTRODUCTION

Small-molecule ligands are generally recognized in protein cavities that are exposed to the surface or easily accessible through low-energy adaptations of a protein (e.g., binding sites in P450 cytochromes). Their binding is experienced via nonbonded interactions established between functional groups of the ligand and amino acid residues exposed to the protein-binding cavity. Upon binding, the ligand has to adopt a particular conformation that allows its functional groups to match with potential counter groups of the protein. Their relative spatial arrangement is defined by the architecture of the binding pocket and the overall fold of the protein. As a rather generally accepted hypothesis in structure-based drug design, pronounced shape complementarity of both interacting partners is assumed as a prerequisite for specific binding.[1] Enzymes are deemed to have optimized their pockets to complement the shape of their substrates or more precisely the transition state of the catalyzed reaction.[2] Therefore, incorporation of transition-state mimetics into potent lead structures is a very popular concept in successful drug development.[3] Nevertheless, it has also been argued that shape complementarity alone is not sufficient to govern molecular recognition, particularly as a comprehensive analysis of shape variations in protein-binding pockets revealed that the pockets tend to be significantly larger than the spatial requirements of the bound ligands.[2] The remaining space is either filled by water molecules or allows the ligands to experience—at least in part—residual mobility in the binding pocket.

It is anticipated that the binding event results in the specific and selective binding of one ligand to one unique target protein, even though recent studies have estimated that a

Additional Supporting Information may be found in the online version of this article.
[†]Deceased 13 January 2011.
Grant sponsors: German Government for the Rehabilitation of Disabled People; Grant sponsor: the Hermann-Leuschner Stiftung.
*Correspondence to: Gerhard Klebe, Institut für Pharmazeutische Chemie, Philipps-Universität Marburg, Marbacher Weg 6, D-35032 Marburg, Germany.
E-mail: klebe@mailer.uni-marburg.de.
Received 13 July 2011; Revised 29 September 2011; Accepted 10 October 2011
Published online 19 October 2011 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.23226

Table I
Cofactors Considered in the Comparative Analysis and Their Splitting into Distinct Fragments of Unique Chemical Building Blocks

Cofactor name	Abbr.	Cofactor fragments used to generate the subcavity
Adenosine diphosphate	ADP	Adenine, ribose, phosphate1, phosphate2
Adenosine triphosphate	ATP	Adenine, ribose, phosphate1, phosphate2, phosphate3
Acetyl CoA	CoA	Adenosine, diphosphate, pantetheine
Flavine adenine dinucleotide	FAD	Adenine, ribose, diphosphate, ribitol, isoalloxazine
Nicotine amide adenine dinucleotide	NAD	Adenine, riboseA, diphosphate, riboseN, nicotinamide
Nicotine amide adenine dinucleotide phosphate	NAP	Adenine, riboseA, diphosphate, riboseN, nicotinamide

xenobiotic experiences on average interaction with at least six protein targets.[4] If binding occurs to other targets aside of the desired one, unwanted cross reactivity might be the consequence. Across a family of related proteins all adopting a similar fold and exposing related amino acids, such cross reactivity can be expected. For example, the potent serine protease inhibitor NAPAP binds either to thrombin and the structurally related trypsin with nanomolar affinity. The crystallographically confirmed bound conformations only differ by a 180° flip of the terminal naphthylsulfonyl group.[5,6] More difficult to predict is the simultaneous binding to proteins of different fold and unrelated functions. In such cases, the binding pocket of the undesirably hit protein exposes residues with similar physicochemical properties but in a distinct spatial arrangement. If the ligand exhibits the required conformational degrees of freedom, it can adopt another conformation and thus achieve potent binding to the unrelated protein. Celecoxib, a potent COX-2 inhibitor, shows unexpected nanomolar binding to carbonic anhydrases, among them CAII.[7] The functionally and fold-topologically unrelated proteins COX-2 and CAII show, nevertheless, distinct similarities in their binding pockets. This fact explains why the COX-2 inhibitor Celecoxib possesses anticancer properties,[8-10] leading to clinically approved reduction of adenomatous colorectal polyps in patients with familial adenomatous polyposis.[11] Other examples of structurally confirmed cross reactivity to remote proteins are, for example, the binding of the anticancer drug imatinib to Abl kinase[12] and oxidoreductase NQO2,[13] a flavoprotein (3fw1), or of the lipid-lowering drug lovastatin to HMGCoA reductase[14] and lymphocyte function-associated antigen (LFA-1), which belongs to the family of beta2-integrins (1cqp).[15]

Usually, the described cross reactivity implies conformational adaptations of a ligand. Therefore, the important question must be asked as to whether a similarly adopted conformation of a ligand is determined primarily (i) by the ligand's force field, (ii) by the shape of the accommodating binding pocket and the exposed physicochemical properties of the recognizing protein residues, or (iii) by the local or overall fold of the target protein next to the binding pocket. Better understanding of this relationship and the relative impact of these contributions to binding will put us in place to predict binding and cross reactivity of therapeutically relevant ligands to proteins. A powerful approach to address such questions is the statistical analysis of structural data; however, a sufficiently large dataset of small-molecule ligands is required that bind repeatedly to proteins of deviating amino acid composition and fold architecture. Even though the PDB[16] now contains more than 70,000 entries, examples as the ones mentioned above are still rather rare. Multiple occurrences of the same ligand binding to structurally distinct proteins have yet only been recorded for cofactors as these molecules are frequently present in proteins and a sufficiently large body of structural information is available. In this study, we therefore analyzed the binding of ADP, ATP, NAD(P)(H), FAD, and acetyl CoA (Table I), which all share an adenosine diphosphate moiety as common substructure.

To classify the properties of a ligand and its hosting protein, we will consider several attributes in a systematic fashion. To avoid analysis of trivial similarities imposed by high sequence homology, we requested the considered proteins, hosting one of the aforementioned cofactors, to exhibit less than 25% sequence identity in the protein chains contributing to the binding pocket. We first studied the conformational properties of the bound ligands. This analysis has been complemented by data retrieved from the Cambridge Structural Database (CSD) containing small-molecule crystal structures.[17] To compare the exposed physicochemical properties across binding pockets, we applied the descriptors developed for Cavbase.[18] The latter approach attributes to binding-site exposed residues up to five so-called pseudocenters that capture the local interaction properties in terms of H-bond donor or acceptor properties and of hydrophobic aromatic (π) or aliphatic interactions. Either, an overall comparison of the entire binding pockets has been performed or, alternatively, a more local approach has been accomplished by splitting the cofactor molecules up to several distinct chemical fragments and using the surrounding subcavities (Table I). The subpockets were subsequently subjected to the formalism of Cavbase to derive descriptors for comparison. Finally, we analyzed the protein architecture next to the ligand-binding sites in terms of preferred folding motifs using DALI.[19] State-of-the-art statistics and clustering techniques have been applied to correlate the extracted structural information.

This article is structured in the following way. The properties of the cofactors in (i) ligand-torsion space, (ii) binding-pocket space (Cavbase), and (iii) protein-folding space are analyzed separately. In each case, a short introduction, also considering related approaches in literature, is given. Then, the analysis strategies, carried out in the three different spaces, are presented followed by a short result section, showing the detected data groupings. Finally, the clustering in the three spaces is mutually faced to detect possible correlations with molecular properties. We also mapped the structural data onto enzyme classification codes (EC numbers[20]), and an affinity score is estimated using the knowledge-based scoring function Drugscore[21] to rank protein–ligand interactions. Most of the results become intuitively obvious by graphical analysis; thus, many of the correlations are presented as images. The observed correlations will be discussed using representative examples; further data analyses can be found in the Supporting Information. The considered dataset and methodological aspects of the analyses are given in the Materials and Methods section.

Analysis of the ligand conformational properties in torsion angle space

Previous studies

A variety of studies have been described in literature to analyze and compare conformational properties of small organic molecules. Of particular interest was the question as to whether a small molecule adopts a different conformation in protein-bound state compared to "less stringent" conditions, for example, in gas phase, in solution, or in a small-molecule crystal structure. These studies come to the conclusion that significant deviations are more the rule than the exception, particularly if the ligand exhibits multiple rotatable bonds.[22] Some comparative studies even attribute the bound conformation to a state energetically well above the global minimum, sometimes distinct from any local minimum.[23,24] Others exemplify the conformational properties in terms of conformational preferences found for local torsion motifs in molecular structures.[25-27] Most crucial in these evaluations is how to relax the geometry found in a protein–ligand complex with physically meaningful restraints to reveal a reasonable geometry to compare with a reference state, for example, in solution. Recently, Butler *et al.* proposed a protocol to produce reasonable energy values that suggest that bioactive conformation and solvated state are not largely different in energy.[28] Some studies have been focused on the analysis of NAD(P)(H), ATP, and FAD.[29-33] Although the earlier works analyzed the ligands in terms of their torsion angle scatter, the more recent studies also relate the conformational properties to the surrounding protein environment using a classification based on the method CATH.[34] Here, the evaluation has been based on visual inspection of molecular superposition followed by an analysis of the matrices of RMS deviations in coordinates and on the comparison of the radii of gyration. Obviously, ligand conformation does not correlate with overall protein fold. In a study of Mitchell,[35] protein sequences have been faced with ligand similarity. Some correlation is suggested when sequence similarity is plotted against averaged Tamimoto ligand similarity.

Data analysis

In our analysis, we concentrated on torsion angles as the most relevant features to describe conformational properties of molecules (Fig. 1). A dataset was compiled from the PDB[16] avoiding sequence homology beyond 25% as described in the Materials and Methods sections. Additional entries from the CSD,[17] the database of small molecule crystal structures, were included (Supporting Information Table S1). To obtain an overview over the scatter of a particular molecular dimension, usually histographical distributions are compared. Crucial for any analysis of discrete data points in terms of histograms is the preselected bin size which partitions a given data space. Because torsion angles are cyclic coordinates, their definition range spans from -180° to $+180^\circ$. To translate the distribution of discrete data points into a continuous density function [Fig. 1(a)], thus avoiding false conclusions from an arbitrary histographical binning, we applied Pareto density estimations (pde-plots, see Materials and Methods section). To reflect the cyclic nature of torsion angles τ (e.g., values at the limits -180° and $+180^\circ$ of the definition range are identical), we expressed their values in terms of sines and cosines and evaluated the Euclidean distances among the individual data points along the unit cycle [Fig. 1(b)]. Finally, torsion angles might be interdependent either by geometrical constraints (e.g., ring closure conditions) or by additional steric restrictions excluding some torsion angle combinations. To detect such torsional intercorrelations, we applied Fisher's exact test [Fig. 1(c)]. It is a significance test to detect correlations by evaluating the contingency or cross-correlation table composed of all torsion angles across the data sample.[36]

Finally, to elucidate a putative clustering of our sample data, we applied self-organizing neural network analysis (Ref. 37, see Materials and Methods section). The input data have been transformed into a set of high-dimensional feature vectors (individual torsion angles) that are subsequently projected onto two dimensions using emergent self-organizing maps [ESOM, Fig. 1(d)]. On such a map, similar vectors fall next to each other, whereas deviating ones end up in separate regions. Because the neighborhood of neurons at the edges of such planar maps would be less dense compared to the center of the map, a two-dimensional toroidal grid is formed for the analy-

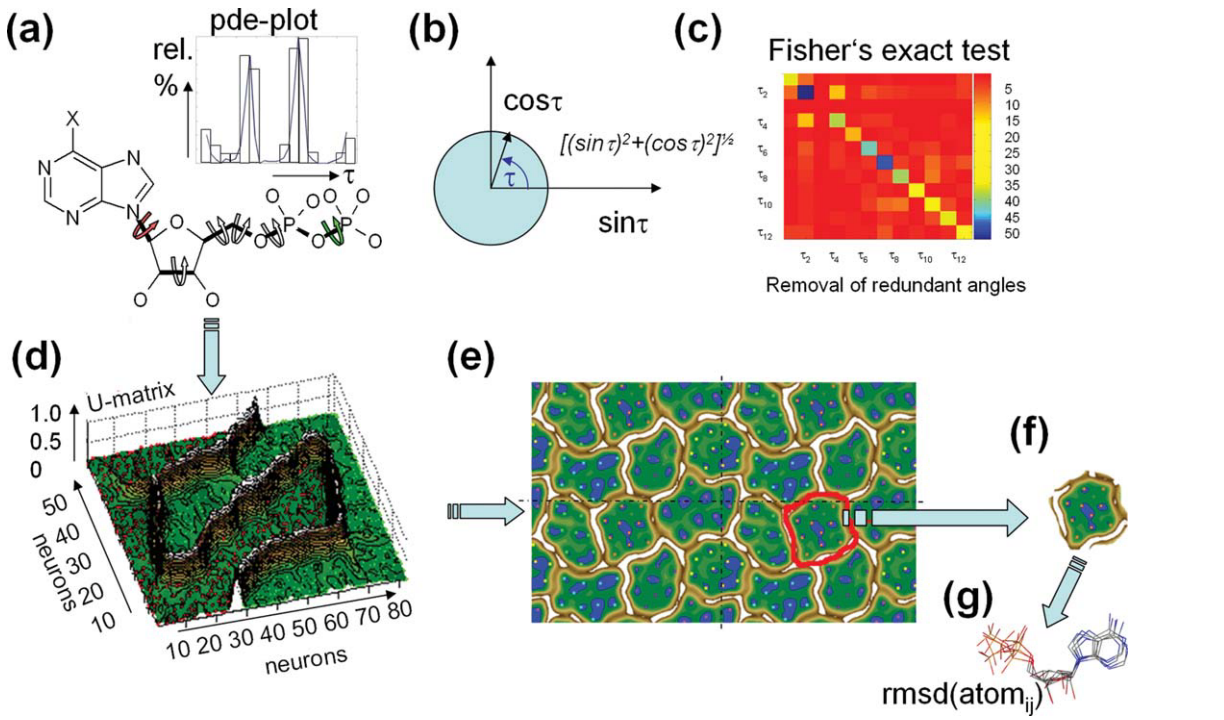

Figure 1

Workflow to analyze data in torsion angle space: (**a**) For a torsion angle fragment to be analyzed, the individual torsion angles τ_x are plotted as histograms and their distribution is studied in terms of Pareto density estimations (pde-plot, see Supporting Information Fig. S17). (**b**) To reflect the cyclic nature of the torsion angles, their values are expressed as sin/cos and evaluated as Euclidian distance along the unit cycle. (**c**) To detect possible intercorrelations among individual torsion angles, Fisher's exact test is performed. (**d**) The preprocessed torsional data are transformed into a feature vector for each entry of the data set. The entire set of feature vectors is analyzed using a projection on a 2D self-organizing neural network. Areas comprising torsion angles assigned to similar feature vectors are visualized by imposing the U-matrix as a kind of "landscape" onto the map. The height in this map corresponds to the distances to neighboring feature vectors. (**e**) The obtained ESOM is displayed with repetitive translational symmetry. The imposed landscape is indicated similarly to geographical maps with valleys (green and blue lakes) and separating mountain ranges (brown with white snow covered summits). Best-matches of the individual entries of the data set are displayed as small squares. (**f**) Data points agglomerating in valleys form individual clusters. (**g**) The individual entries in a common cluster were superimposed using the atomic coordinates of the fragments under consideration. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

sis where the edges of the grid are mutually connected to avoid any border effects.[38] For subsequent visualization, the toroidal maps are displayed as planar maps with repetitive translational symmetry. This means neurons on the top are neighbored by the neurons on the bottom and neurons on the left by those on the right (like a "Pacman universe"). To detect putative clustering among the feature vectors on the ESOM, a kind of "3D landscape of valleys and mountains" is used based on values of the corresponding U-matrix. The U-height at each neuron is a sum of distances to the neighboring feature vectors and is displayed as a color-coded contour plot on top of the two-dimensional projection, very similar to geographical maps. Data points clustering in a valley between mountain ranges represent similar feature vectors and thus similar torsion angles. Clearly separated and distinct clusters will assemble in detached valleys. Entries distinct from any other examples in the data set (singletons) will end up in deep wells separated by high mountain ranges (in eye-catching manner "covered by snow") of the U-matrix [Fig. 1(e)]. Entries on such maps (so-called best-matches[39]) can be colored according to an additional property, for example, the chemical compound class they originate from or a particular torsion angle found in the individual entries. It must be pointed out that any existing clustering across the data sample becomes obvious without any predefined or anticipated knowledge about the expected number of clusters. Finally, all ligands attributed to one conformational cluster were mutually superimposed using an RMSD superposition based on the matching atoms in the ligands [Fig. 1(g)].

Results in torsion angle space

Adenosine diphosphate moiety (adenosyl fragment)

We started our analysis investigating the adenosine diphosphate moiety (named "adenosyl fragment" in the following), the largest substructure shared in common by

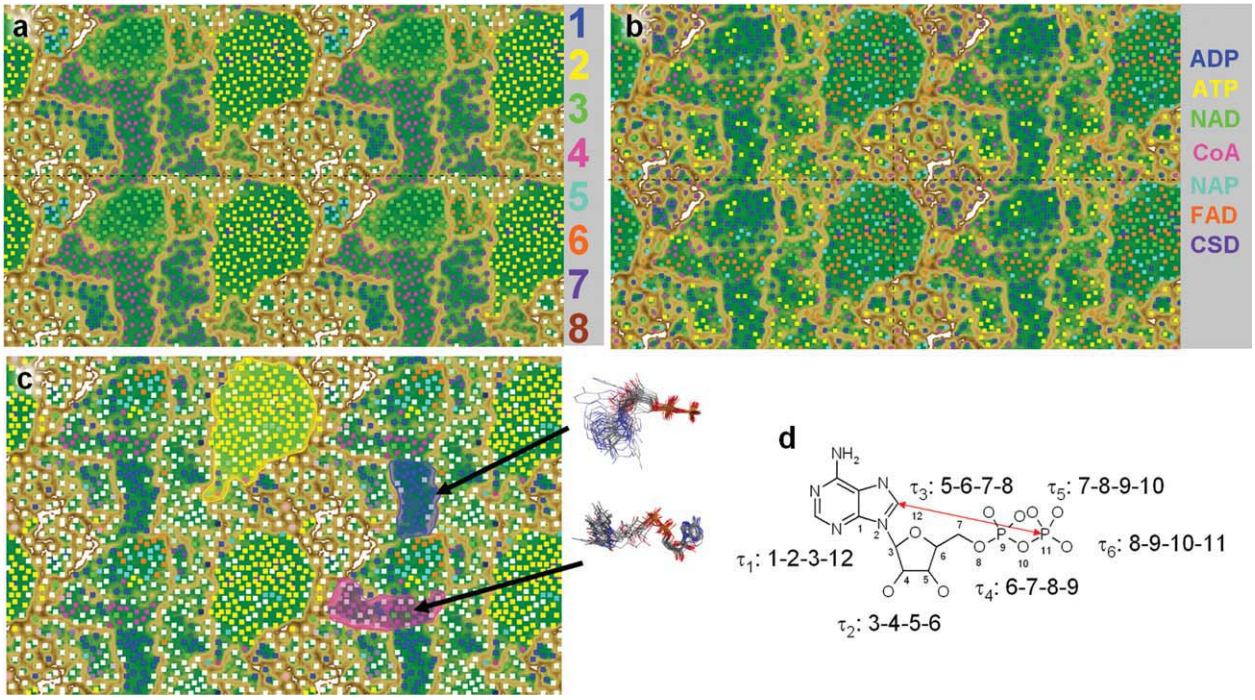

Figure 2

Chemical composition, assignment of torsion angles, and ESOM-torsional map for the adenosyl diphosphate moiety (named "adenosyl fragment"). The ESOM (**a**–**c**) is superimposed with the U-matrix to analyze the torsion angles τ_1–τ_6 of the adenosyl fragment. Best-matches of the individual torsion fragments in the data set (487 entries) are indicated by small squares. The map shows eight larger clusters corresponding to distinct conformational families, which agglomerate in deep valleys (green), and a large number of singletons separated by pronounced mountain ranges (brown, partially covered with "snow"). (a) In the ESOM map, the individual entries are color-coded according to their membership to one of the individual clusters. Singletons are colored in white. (b) The best-matches on the ESOM have been colored according to the parent cofactor molecules from which they were extracted (NAD = NAD(H) and NAP = NADP(H)). (c) In this representation of the ESOM, the best-matches have been color-coded according to their membership to clusters in the cavity space (cf. Fig. 6). (**d**) Assignment of the individual torsion angles τ_1–τ_6 of the adenosyl fragment, in addition, the distance between C2 of the adenine moiety and the terminal phosphorus atom (red arrow) has been evaluated (cf. Fig. 3). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

all considered cofactors. The relevant torsion angles were extracted according to the atom numbering shown in Figure 2(d). A total of 459 entries have been retrieved, originating from the six cofactors (Supporting Information Table S1). This data sample was supplemented by 28 entries extracted from small-molecule crystal structures in the CSD (Supporting Information Table S2). A set of six torsion angles was considered in the analysis. As obvious from the different pde-plots, the torsion angle τ_1 about the *N*-glycosidic bond to the adenine moiety populates, either in PDB or in CSD, in the vast majority of cases around -90° (*anti*), whereas a small portion shows *syn* orientation. The ribose ring has already been studied in detail for its conformational preferences.[29,40-42] In agreement with these results, we find two mainly populated conformers with C2′-*endo* and C3′-*endo* puckering of the ribose ring. They show a population ratio of about 2/3 to 1/3 (depending on the studied subset). Because of internal redundancies of the cyclic ring portion, we selected only one ring torsion angle (τ_2) to describe the adopted ring pucker. The remaining four angles τ_3–τ_6 record the properties of the attached methylene pyrophosphate chain. According to the pde-plots, they occur with different distribution, mostly in agreement with +/- *gauche* and *anti* orientation; τ_4 shows only monomodal distribution around 180°.[29,30]

In Figure 2, the ESOM obtained for the adenosyl fragment is shown. Data scatter can be grouped into eight clusters; the population of the individual cluster is apparent from the mapping of best-matches onto the ESOM map (a "best-match" is the representation of an original data point on the trained ESOM map).[39] Furthermore, the chemical types of the cofactors have also been mapped onto the ESOM. The largest cluster 2 (yellow) is mainly populated by NAD(P)(H) and FAD entries, whereas the clusters 1, 3, and 4 are more strongly dominated by ADP and ATP. Interestingly, the cluster 5 and several entries distinct from the remaining ones next to the latter cluster display adenosine diphosphate geometries found only in small-molecule crystal structures. Remarkably, the data in the small-molecule structures depart significantly from those in protein complexes (see below).

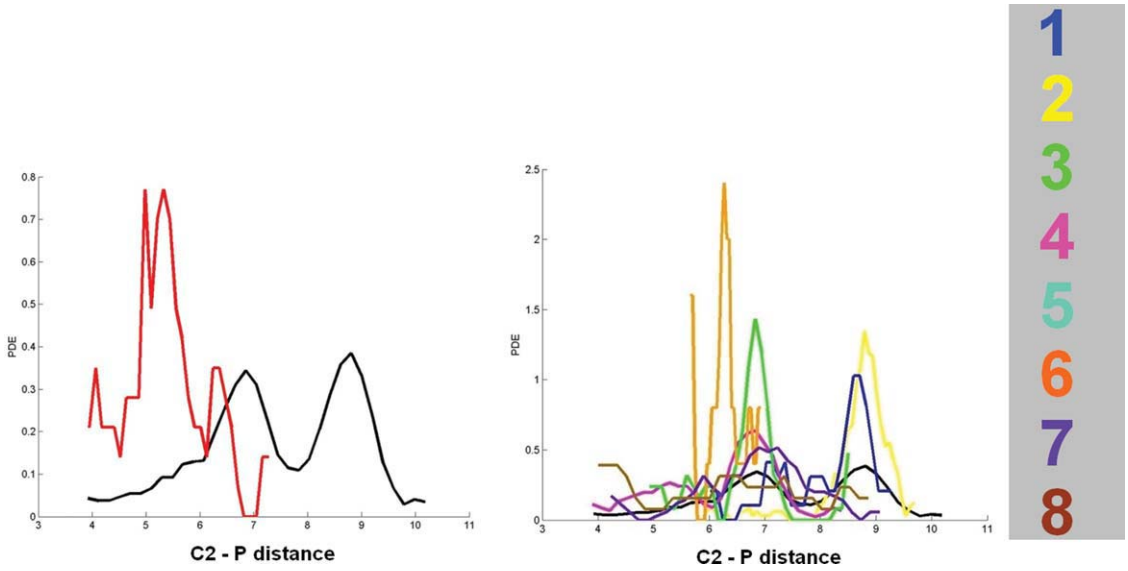

Figure 3

Pareto density estimation (pde-plot) for the distance *d* between adenine C2 and terminal phosphate group in the considered adenosyl fragment. The entries from the CSD (left, red) show significantly shorter distances than the examples from the PDB. The latter exhibits a bimodal distribution. Factorizing the PDB data, according to the clustering detected by the neural network analysis, shows that clusters 1 and 2 exhibit long distances, whereas short ones are observed for clusters 3, 4, and 7. The former ones are mainly composed by NAD(P)(H), FAD, and CoA, whereas the latter ones contain more ADP and ATP entries. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

As mentioned, the *N*-glycosidic bond occurs overwhelmingly with *anti* geometry, entries with *syn* orientation end up in regions occupied by singletons (Supporting Information Fig. S1). The C2′- and C3′-*endo* conformation of the ribose ring separates well among the clusters, only cluster 1 hosts examples with both conformations. Torsion angles τ_3 and τ_6 occur nearly exclusively with one value per cluster, whereas τ_5 spreads across the entries in clusters 1, 3, 4, and 7 over several conformations (see Supporting Information Fig. S1). To obtain a better insight what makes the entries in the small-molecule crystal structures distinct from those in the protein complexes, we determined the distance between C2 of the adenine moiety and the second phosphorus atom of the pyrophosphate unit [see Fig. 2(d), red arrow]. The pde-plot (Fig. 3) for this distance recorded between 4 and 10 Å shows much shorter distances in CSD data (red) than in protein entries (black).

ADP and ATP

Subsequently to this overview over the adenosyl fragment, we focused on the individual cofactors. At first, we considered ADP (125 entries), regarding only cases with ADP as complete ligand. Six torsion angles have been considered in the analysis and they all prove to be independent. One dominantly populated cluster is found (66 entries, Supporting Information Fig. S2) apart from several additional, but weakly populated ones. Furthermore, many entries end up on mountain ranges indicating low conformational similarity with any of the other entries. The *N*-glycosidic bond occurs predominantly in *anti* orientation and the ring pucker splits almost equally between the two preferred conformers.

In the case of ATP (79 entries, Supporting Information Fig. S3), an even larger scatter over individual conformers is observed. Only one notable cluster with eight entries can be assigned. Already the maps for the adenosyl fragment indicated a less systematic distribution for ATP than ADP. In Figure 2(b), the ATP examples (yellow) are more scattered across the map than the ADP ligands (blue). Aside from three exceptions (PDB codes: 2e5y, 1b8a, and 2j3m), which show the terminal end of the phosphate unit close to the adenine moiety, the ligand preferentially adopts an extended conformation. The *N*-glycosidic bond is in most cases in *anti* orientation (apart from nine entries) and the C2′-*endo* conformer is slightly more populated.

FAD, NAD(P)(H), and CoA

For the two larger cofactors FAD (Supporting Information Fig. S4) and NAD(P)(H) (Supporting Information Fig. S5), pronounced conformational clustering is observed. For FAD (91 entries), one dominant conformer ($\sim$ 40% of the data) is found and a second much smaller cluster. Apart from these, the other FADs scatter across the ESOM and indicate a large number of conforma-

tional singletons. No entry with *syn* orientation about the *N*-glycosidic bond is found and, apart from seven entries, all FADs adopt an extended conformation.

The oxidoreductase cofactor NAD(P)(H) (122 entries) shows two significantly populated clusters (16 and 30%). The remaining examples end up as individual singletons. No particular preference is observed for either the phosphorylated or the unphosphorylated derivatives. The entries in the two highly populated clusters differ by the orientation of the carboxamide group in the nicotine amide fragment. The *N*-glycosidic bond occurs mostly in *anti* orientation. Some examples with an internal stacking of the adenine and nicotine amide fragment are observed (e.g., 2d38, 2bkj, and 2d37).

Finally, we looked at the cofactor acetyl CoA (42 entries), the by far largest ligand considered in this analysis (Supporting Information Fig. S6). ESOM training, based on all 19 nonredundant torsion angles, could not detect any meaningful clustering. Subsequently, we only considered the pantetheine moiety (thus excluding the 3′-phospho-adenosine diphosphate portion) in the analysis. Interestingly, for the remaining nine angles, we observed one distinct cluster with a preferred conformation along the pantetheine portion. Nevertheless, also for this portion, a large number of singletons is observed. The C2′-*endo* conformer is preferred and about 21% of all entries show *syn* orientation of the *N*-glycosidic bond.

Analysis of protein-exposed physicochemical properties in the surrounding binding pockets

Previous work

A large number of studies have been dedicated to the determination, analysis, and comparison of binding sites. They either detect cavities or clefts at protein surfaces and match them to one another in terms of geometry, exposed features, algorithmically derived properties, sequence similarities, or spatial arrangements of cavity residues.[43-48] Another large group of approaches tries to identify and compare specific geometrical patterns of conserved amino acids or secondary structure elements to relate and cluster proteins. The latter methods will be summarized with the introduction of the analysis in fold space (see below). With respect to the binding-site detection, some algorithms have been developed that use phylogenetic information to predict the location and properties of such sites in proteins. All methods have in common that the complex structure of a binding pocket is translated into a simplified model representation that captures the most important features for making binding sites different but well suited for a comparative analysis.

Kahraman *et al.* developed a method for shape comparison of binding pockets including bound ligands using a description in terms of spherical harmonics.[2] Baroni *et al.*[49] compared protein-binding sites using fingerprints of ligands and proteins, which are based on interaction properties computed by chemical interaction probes. A similar analysis has been performed by Henrich *et al.*[50] based on molecular interaction fields. Also, properties such as the electrostatic potential on protein surfaces can be used for comparative analyses.[51] A number of approaches apply graph-based comparative algorithms. Najmanovich *et al.*[52] detected binding pockets by Surfnet[53] and classified exposed types of atoms for a comparative analysis using a clique-search approach. Park and Kim[54] used the information about bound ligands and their mutual similarities to detect relationships between protein-binding sites. The approach of Jambon *et al.*[55] uses graphs of triangles of chemical groups to detect local similarities between proteins via a graph-matching algorithm. The program CPASS of Powers *et al.*[56] compares ligand-defined active sites to determine sequence and structural similarity without maintaining sequence connectivity. Other approaches evaluate solvent-accessible protein surfaces for comparison[57]. In another similarity analysis, topological and physicochemical properties have been projected from cavity-lining protein residues to a sphere inscribed into a protein-binding pocket.[58] Another set of algorithms evaluates the properties responsible for the recognition of ligands at binding sites. One prototype of this category is Cavbase, a method introduced by us.[18] It attributes pseudocenters to represent the location and physicochemical properties of the binding-site exposed residues. Similar concepts have been implemented into algorithms such as Site Engine.[59] The Cavbase approach is described in the following section as this concept has been applied in this study to analyze the cofactor-binding sites in cavity space.

Data analysis

In this study, we wanted to describe the binding pockets, hosting one of the cofactors, in terms of exposed physicochemical properties. For this purpose, attributes directly linked to the amino acids can be evaluated, but also derived properties such as the electrostatic potential, sequential motifs, or hydropathy indices can been exploited. In our evaluation, we decided to apply the rule-based descriptors developed for the Cavbase approach.[18] This method serves as a fast engine to mutually compare binding pockets across a large number of proteins. In Cavbase, a protein is embedded in a regularly spaced grid using the program LIGSITE.[60] Any lattice intersections coinciding with protein atoms are discarded. The remaining lattice points are scored for their burial in surface depressions and adjacent points are clustered together to reveal contiguous cavities. The cavity surface is approximated by the grid points in contact with protein atoms. The atomic coordinates of the residues flanking the thus defined cavities are reduced to a

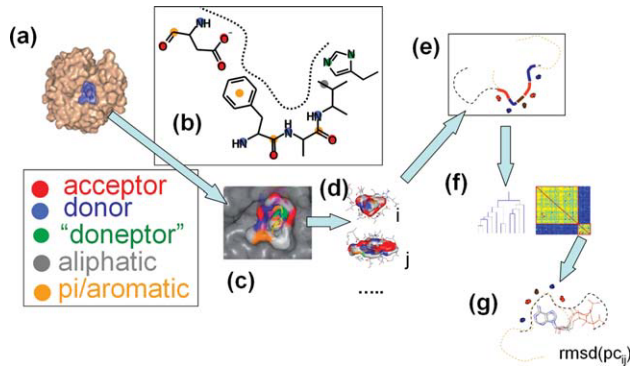

Figure 4

Workflow to analyze data in binding cavity space: (**a**) The analysis of the physicochemical properties of the binding pocket hosting a particular ligand or ligand portion starts with the extraction of a binding pocket (blue). (**b**) The functional groups of the amino acid residues flanking the binding site are classified by well-placed pseudocenters (pc) that describe H-bond donor, acceptor, or mix donor/acceptor (doneptor) properties. Furthermore, hydrophobic properties are assigned. (**c**) In addition, surface patches next to the allocated pseudocenters are attributed. (**d**) In the following step, all extracted binding pockets are mutually compared. (**e**) A clique detection algorithm is applied to find common subgraphs among the assigned pseudocenters. Once a shared subgraph is found a mutual alignment of the cavities is calculated and a similarity score is calculated also considering the matched surface patches shared in common. (**f**) The computed similarity indices are used for clustering and a similarity matrix is produced. (**g**) Within indicated clusters the ligands are superimposed using the transformation matrix determined by the mutual pseudocenter match. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

set of preannotated pseudocenters (Fig. 4). They represent, based on a given amino acid geometry, the essential physicochemical features that determine molecular recognition of a putative bound ligand [Fig. 4(b)]. In total, five properties are considered, relating to H-bond donor, H-bond acceptor, or simultaneous donor/acceptor (doneptor) facilities and hydrophobic lipophilic and aromatic contacts. In addition to the assignment of the pseudocenters, Cavbase evaluates the surface patches that fall next to allocated pseudocenters [Fig. 4(c)]. The comparative binding pocket analysis matches sets of pseudocenters shared by two proteins and computes a mutual superposition of the two binding sites; thus, no information about the spatial position of the atoms of the bound ligands is used to compute the protein superposition [Fig. 4(e)]. Different scoring schemes have been defined to evaluate the degree of similarity achieved. The similarity scores obtained for the mutual comparison of all considered binding pockets serve as input for different clustering algorithms. Before clustering, the calculated similarity scores have been normalized by dividing the achieved mutual score with the arithmetic mean of the self-fit of the individual binding pockets (see Materials and Methods section).

In the following analysis, we applied the Cavbase comparison scheme to two sets of predefined protein cavities. We started with all entries considered in the torsion space analysis; however, we required the cofactors to be fully embedded in the protein-binding pocket as extracted by the LIGSITE algorithm[60] (see Materials and Methods section). These pockets will be named "surround" pockets in the following (Fig. 5). To perform a more local similarity analysis, we also split the different ligands, considered in this study, into several distinct fragments by chemical means (Table I). For example, an ADP ligand (Fig. 5) has been fragmented into an adenine moiety, its ribose portion, and the two distinct phosphate groups. Subsequently, those parts of the surround pocket that fell within 4 Å of any atom of the generated fragments were classified adenine, ribose, or phosphate subpocket. This splitting was applied either to the pseudocenters and the corresponding surface patches.

According to the achieved subpocket definition, a comparable pocket analysis has been performed by evaluating the similarity of the subpocket descriptors in terms of the Cavbase methodology. Clustering of the normalized Cavbase similarity indices has been performed by an agglomerative hierarchical clustering using Ward's criterion (see Materials and Methods section). A dendrogram

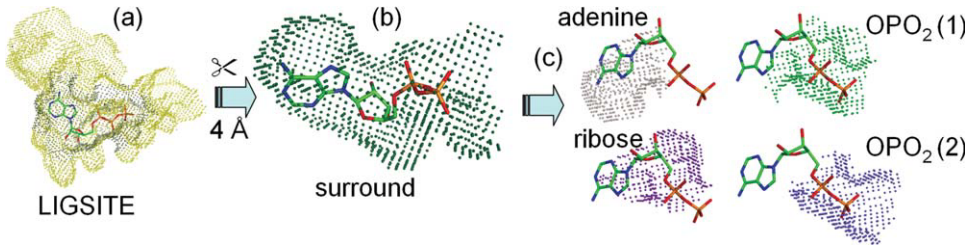

Figure 5

The surround pocket is extracted from the original LIGSITE pocket stored in Cavbase (**a**). It is defined to enclose the entire ligand and comprises all surface points and pseudocenters that approach any of the ligand atoms to closer than 4 Å (**b**). To perform a local analysis, the ligand has been fragmented into different chemical building blocks (**c**). In the present example, ADP is partitioned into an adenine, ribose, and two phosphate groups. Subsequently, the subpocket around the split ligand portions is calculated similarly as for the surround pocket. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

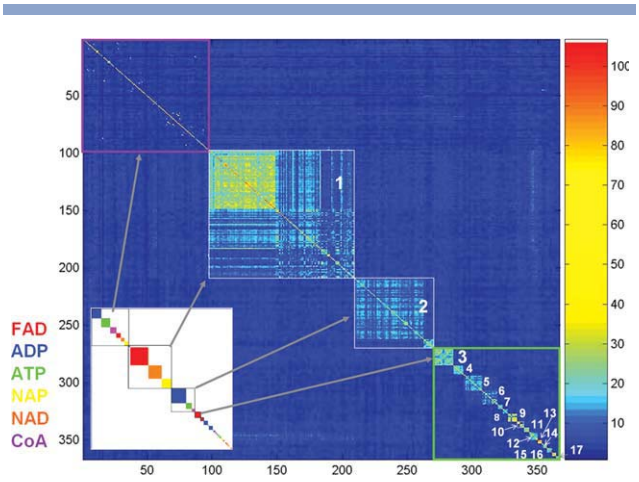

Figure 6

Based on 367 cofactor-surround pockets, a similarity analysis based on the Cavbase descriptors has been performed. The diagram shows the pair-wise similarity matrix, and the mutual similarities are indicated by a color-coding from 0 to 100% (low similarity blue, high similarity red). In total, 17 distinct clusters are apparent from the analysis. In the upper left part of the diagram (magenta box), structures are listed that show only little similarity with any other entry. On the bottom left, an inset is shown that gives the assignment of the different cofactor chemotypes (in a color-coding) to the 17 clusters. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

[Fig. 4(f)] for the obtained clustering of either the surround or the subpockets has been generated. The corresponding similarity matrix is analyzed by superimposing a color-coding scheme to indicate the degree of achieved similarity. Finally, for each detected cluster, the corresponding cofactors have been mutually superimposed by applying the transformation matrix defined by the matching pseudocenters [Fig. 4(g)].

Results in the surrounding binding pocket space

All cofactors

At first, we calculated Cavbase similarity scores across all 367 fully embedded and mutually aligned surround pockets hosting one of the considered cofactors (Supporting Information Table S2). After clustering, the similarity matrix shown in Figure 6 has been obtained. Two major clusters (Fig. 6, 1, 2) are holding 112 and 61 entries. Furthermore, a set of 14 smaller clusters (3–17, green box) with 3 to 15 entries is detected. Apart from these obvious clusters, 97 entries (upper part, magenta box) display predominantly singletons (named cluster 0, cf. discussion). Considering solely Cavbase pseudocenters that originate from molecular portions composing the protein backbone, we obtained a more clear-cut clustering than producing a clustering based on pseudocenters residing on amino acid side chains only (Supporting Information Fig. S7). It is remarkable to see that a stronger discrimination of the different pockets is obtained by the properties originating from the backbone than from the side-chain atoms. Mapping the six cofactor chemotypes onto these clusters indicates a well-structured distribution. Many clusters are predominantly populated by one cofactor type only, the two large clusters (1, 2) host nearly exclusively either FAD and NAD(P)(H) or ADP, ATP, and FAD (Fig. 6, lower left part).

ADP

Subsequently, we concentrated on the individual cofactors. Thus, the aforementioned clusters were assembled in a way to consider cavities that host cofactors of the same type only (Supporting Information Table S3). At first, complexes with ADP as cofactor were analyzed, regarding 95 ADP surround pockets. The corresponding dendrogram (Supporting Information Fig. S8) suggests five clusters of which two can be further partitioned (Supporting Information Fig. S8, 1, 2 and 5, 5a) into smaller ones populated by three or four entries. The large cluster 2 remains rather heterogeneous. Superposition of the accommodated ligands according to the matched pseudocenters in clusters 1, 3, 4, 5, and 6 shows that different regions of the ligands are oriented in similar fashion in the pockets (Fig. 7). Particularly, the large cluster 5 with 39 entries shows high similarity only in the region of the pyrophosphate group. Interestingly enough, the subsequently performed analysis decomposing the surround pocket into four subpockets hosting an adenine, ribose, and the two phosphate fragments (Table I) renders the clusters 1, 3, 4, and 6 prominent regarding the adenine subpocket only. In contrast, the large cluster 5 is only well returned if the descriptors of the subpockets hosting the pyrophosphate portions are used. This analysis clearly shows that high Cavbase similarity is only achieved if the ligand exhibits a common orientation in the corresponding pockets.

ATP

Following a similar protocol, the pockets hosting ATP ligands were analyzed. A total of 53 ATP-cofactor-surround cavities decompose into five different clusters (dendrogram Supporting Information Fig. S9). One cluster (2) remains internally rather diverse. For the remaining entries, the similarity matrix of Cavbase indices shows one major (cluster 1) and three smaller clusters (3–5, Supporting Information Fig. S10). Similarly to ADP, here the large cluster 1 comprises entries that exhibit similar orientation along the oligophosphate group. Accordingly, only clustering of the phosphate subcavities renders this group of ligands prominent, whereas the three other clusters become already apparent considering in particular the adenine subcavity. In cluster 3, the conformational and orientational similarity across all matched ligands is rather high. In consequence, all five

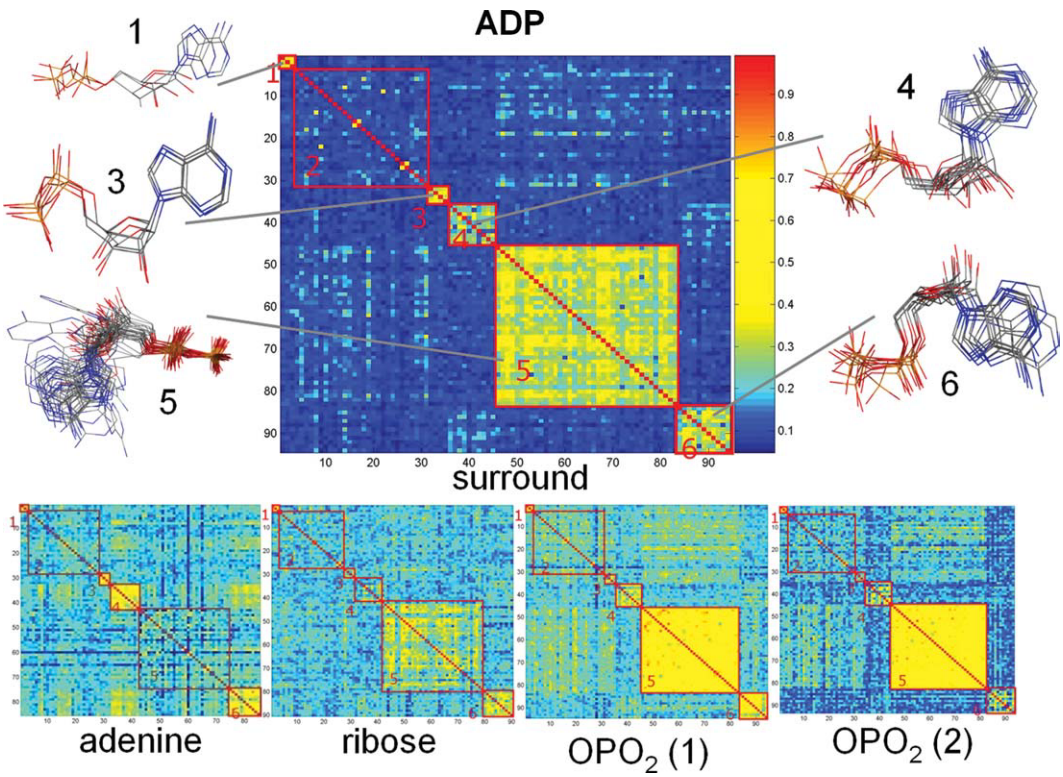

Figure 7

Clustering of Cavbase descriptors in binding pockets hosting ADP. Considering the ADP-surround pockets, five clusters with pronounced mutual similarity are indicated. The superposition of the corresponding bound cofactor molecules is shown and their alignment has been calculated using the matched pseudocenters. Below, the same analysis is shown, now considering the four subcavities. Cluster 4 is strongly indicated by the adenine subpocket, whereas cluster 5 becomes only transparent once binding to the phosphate subpockets is analyzed. This correlates well with the fact that common orientation is only found for these ligands along the diphosphate moiety. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

subcavities produce a convincing accentuation of this cluster. In case of clusters 4 and 5, the triphosphate group scatters more strongly. This parallels to a rather fuzzy and less-well detectable clustering of the subpockets hosting the three phosphate groups.

FAD

In Figure 8, the analysis of the surround pockets hosting FAD as cofactor (91 entries) is shown. Three distinct clusters with high internal similarity are indicated, a fourth cluster remains quite heterogeneous. In clusters 1 and 2, almost all the fragments with the exception of the isoalloxazine moiety align quite well. In agreement with these findings, the analysis of the individual subpockets accentuates these two clusters comparatively well, whereas the cavities surrounding the isoalloxazine fragment barely form a common cluster. In contrast, the third cluster comprises pockets that show the isoalloxazine moiety in common orientation. Particularly here, however, the ribose and adenine portion show widely scattered arrangements. Accordingly, a clustering based on the subpockets hosting these cofactor fragments does not indicate cluster 3 in the analysis. Finally, we mapped the Cavbase clustering observed with respect to the surround pockets onto the torsion angle map obtained for this cofactor (Fig. 8, lower part). It is interesting to note that the entries corresponding to the major cluster 2 in cavity space also fall into the largest cluster in torsion angle space (best-matches indicated in magenta). Only some entries in the latter space, assigned to singletons, are also members of the large Cavbase cluster. The members of the remaining clusters 1 and 3 in cavity space correspond in the torsion angle analysis to entries (best-matches in black and green) that fall into mountain ranges, thus describing conformationally rather distinct examples.

CoA

A similar analysis of the CoA cofactors (32 entries) exhibits only two distinct clusters with pronounced internal similarity (Supporting Information Fig. S11). Considering the mutual alignment of the ligands hosted by these cavities based on the matched pseudocenters shows that ligands corresponding to the first cluster exhibit common orientation only along the pantetheine moiety, whereas

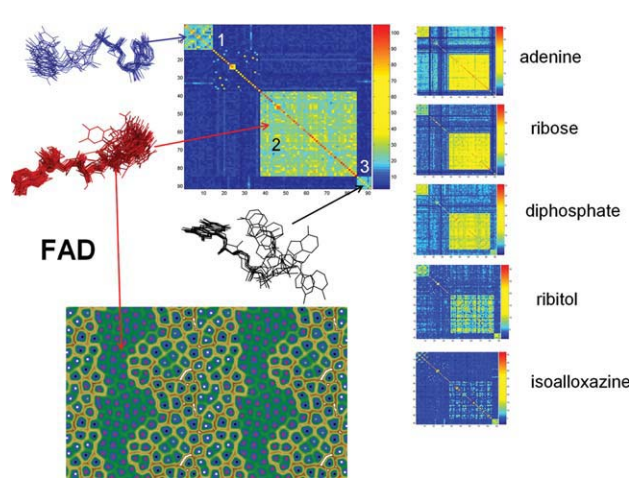

Figure 8

Clustering of Cavbase descriptors in binding pockets hosting FAD. Considering the FAD-surround pockets, three clusters with high mutual similarity are indicated. The superposition of the corresponding bound cofactor molecules is shown and their alignment has been calculated using the matched pseudocenters. On the right, the same analysis is shown, now considering five subcavities. Clusters 1 and 2 are strongly indicated by the adenine, ribose, and diphosphate subpockets, whereas cluster 3 becomes only transparent once binding of the pocket hosting the isoalloxazine ring is studied. This correlates well with the observation that a common orientation is only found for FAD with respect to the tricyclic moiety (black). On the lower left, a mapping of the three Cavbase clusters onto the distribution in torsion angle space is shown. Although entries from cluster 2 in cavity space coincide with the predominant torsion angle family, entries from clusters 1 and 3 in cavity space map to singletons in torsion space. This indicates pronounced conformational distribution in the latter two cavity clusters. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

for the second cluster also the diphosphate group shows common alignment of the hosting ligands. Performing an analysis of the corresponding subpockets reveals a very similar picture as for the previously described cofactors. Again, the analysis using the pantetheine subcavity returns the two clusters similar to the ones detected by the analysis of the surround pockets, whereas the characterization of the diphosphate subcavity only renders the major cluster found by the study of the surround pockets as sole agglomeration. Finally, when the adenine subcavity is used for the analysis, no cluster of similar structures becomes apparent for the CoA data set.

NAD(P)(H)

The analysis of surround binding pockets (110 entries) hosting the cofactors NAD(P)(H) reveals one predominant cluster (Supporting Information Fig. S12, 5), five small and internally rather homogeneous clusters (1, 3, 4, 6, and 7), and one agglomerate (2) with rather diverse entries. The major cluster 5 shows some structuring in two families, which can be attributed to either NAD(H) (54 entries) or NADP(H) (46 entries) as cofactors. This splitting becomes even more apparent when considering the clustering of the adenine subpocket only. The small clusters 1 and 7 host ligands with very similar conformations. This observation in torsion angle space matches well with the high similarity of the surround pockets or the individual subpockets. Obviously, across all entries with high similarity in the adopted conformation also a high Cavbase similarity is given.

Analysis of commonly found folding motifs around the cofactor-binding sites

Previous work

A large body of computational approaches has been suggested to represent and compare proteins in terms of local amino acid architecture. Particularly for enzymes, a small set of functionally important residues tends to be preserved with unchanged relative geometry across even remotely related proteins. The Catalytic Site Atlas database[61] contains a catalogue of structural templates of amino acids derived from the catalytically important residues in enzymes. The methods pvSOAR[62] and CASTp[63] consider residue data in pocket and void surfaces also evaluating evolutionary information to predict their location. A number of algorithms have been suggested like SPASM, RIGOR, or Jess[64,65] or the methods developed by Besl and McKay,[66] Nussinov and Wolfson,[67] or Stark and Russell[68,69] to scan proteins for the occurrence of such amino acid template motifs in comparative analyses.

Another set of approaches uses clique-type algorithms based on coordinate data of binding-site residues (e.g., CSC,[70] C_α atoms or triplets of exposed atoms, e.g., TESS[71,72]). Some include information about side-chain orientations (ASSAM[73,74]) or perform searches for commonly shared residue geometries (Query3d[75] or DRESPAT[76]). In several of these approaches, genetic algorithms are applied to detect a given protein similarity.[77-79] Holm and Sander developed the program DALI to detect similarities among proteins based on the spatial arrangement of secondary structure elements.[19] This method can be seen as a prototype for the analysis of folding patterns in proteins based on 3D coordinates. It has been applied in this study and its algorithms will be described shortly in the next section.

Data analysis

To relate the cavity analysis of exposed physicochemical properties, described in the previous section, with the local folding architecture of the proteins hosting one of the six considered cofactors, we performed an analysis using the program DALI.[19] The algorithm in this program calculates optimal pair-wise spatial alignments of protein structures. Using the 3D coordinates of each protein, residue-by-residue (C_α–C_α) distance matrices are

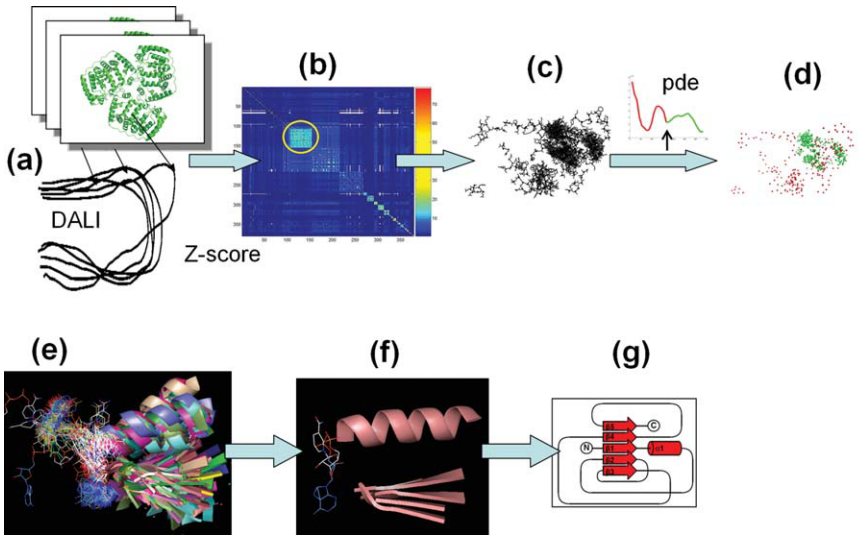

Figure 9

Workflow to analyze data in fold space: (**a**) To detect similarity in folding architecture of the analyzed cofactor complexes, amino acid stretches with similar fold are detected and extracted by DALI. (**b**) The computed *Z*-scores are mapped onto the similarity matrix obtained by the Cavbase analysis. (**c**) To find the consensus folding pattern next to the cofactor-binding pockets in the different clusters the most frequently matched amino acids in the pair-wise comparisons are extracted and aligned in space according to the mutual superposition suggested by the matched pseudocenters in Cavbase. (**d**) To extract the shared motifs, Pareto density estimations were consulted to find those C_α atoms that were most frequently matched by the entries in one cluster. (**e**) They gave rise which amino acids were included in the consensus motif and displayed by a molecular graphics tool. (**f**) From this superposition one archetypic entry is shown. (**g**) Using the program TOPS, a schematic sketch of the fold topology is given (see Fig. 10). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

determined [Fig. 9(a)]. After first decomposing these matrices into elementary contact patterns, for example, of hexapeptide–hexapeptide submatrices, similar contact patterns in the two matrices are paired and combined into larger consistent sets of pairs. A similarity score is optimized to find the largest sets of equivalent intramolecular distances. The program produces multiple alignments that are scored as best, second-best, and so forth solutions. It generates for each pair-wise comparison a spatial transformation (rotation matrix and translation vector) to mutually superimpose the aligned structures.[19] To discriminate among the differently ranked solutions in our application, we also applied the transformation to the coordinates of the bound cofactors. We subsequently selected as best-scored solution for our analysis the one which showed the shortest Euclidian distance among the centers of the bound cofactors. The individual scores, as computed by DALI for the finally selected alignments, were subsequently used to construct a similarity matrix analogous as in case of the cavity space analysis. This matrix was subjected to a cluster analysis as described above for the cavity space [Fig. 9(b)]. For all clusters found in the cavity comparison, we detected corresponding similarities in the local folding space. This also supports the observation that the Cavbase analysis based on backbone atoms only received the better clustering performance (Supporting Information Fig. S7).

To extract the commonly shared fold motifs next to the cofactor-binding sites in the different clusters, we closely inspected their spatial orientations by computer graphics. This visual inspection of the peptide stretches, suggested by DALI, clearly indicated commonly shared secondary structural motifs; however, they showed significant spatial deviations or differences in lengths and orientations. This fact made a straightforward RMSD superposition based on C_α atom coordinates to map consensus fold motifs rather difficult. To circumvent this problem, we started for each individual cluster with the amino acid stretches detected by DALI. We then selected those amino acids that were matched in all pair-wise alignments above a given threshold. To adjust this threshold, we consulted Pareto density estimations (see Materials and Methods section). On the basis of these amino acids, we obtained the superposition shown in Figure 9(c), which exhibits high density of amino acids in some areas. To extract only those C_α atoms that fell into regions of high density with neighboring C_α atoms [Fig. 9(d), green], we applied again a Pareto density estimation using a threshold level of 30%. On the basis of the remaining C_α atoms, we visualized the full secondary structures with PyMol[80] [Fig. 9(e)]. From these, we selected as representative structure the one that fell closest to the geometrical average of all entries [Fig. 9(f)]. Using the program TOPS,[81,82] we generated a schematic

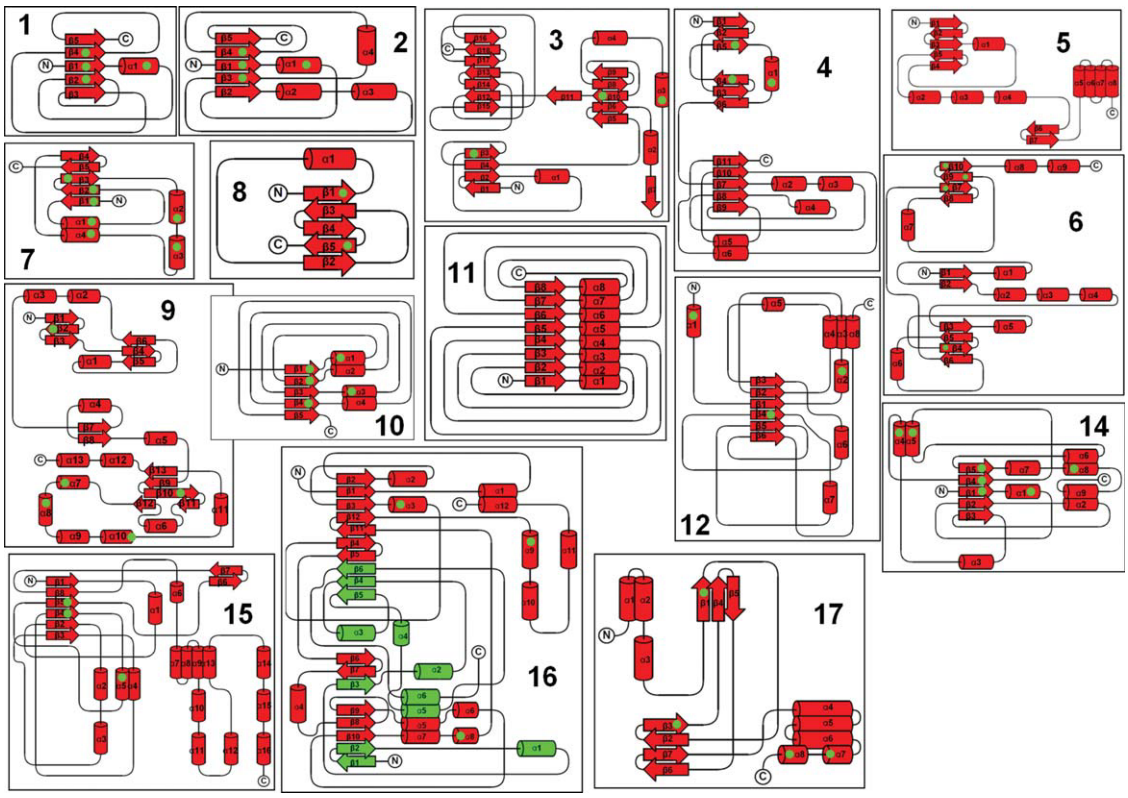

Figure 10

Using one representative entry from each of the different clusters found in fold space, a schematic sketch of the topology of the detected fold motifs is generated using the program TOPS; arrows indicate β-strands, cylinders stand for α-helical portions. The listed numbers correspond to the labels assigned in Figure 6 to distinguish the different clusters found in cavity and fold space. In cluster 16, two polymer strands compose the pocket, and their secondary structure elements are depicted in red/green. Across the individual fold patterns the secondary structure elements participating directly in cofactor binding are marked by a green dot. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

drawing of the local fold topology. The obtained fold sketches are summarized in Figure 10 for the shared local fold motifs.

Results for commonly shared folding motifs

Using the above-described procedure, we detected that repeatedly the same local folding motifs are found next to the matched Cavbase cavities. A virtually identical clustering is achieved in cavity and folding space [cf. Figs. 6, 9(b) and Supporting Information Fig. S7]. The extracted local fold motifs show a broad range of distinct topologies (Fig. 10). By mapping the six cofactor chemotypes onto these fold motifs, it becomes evident which cofactors are recognized by which local folds.

In the following, selected clusters will be described in detail. For the largest cluster 1 (112 entries, assignment see Fig. 6), a motif formed by five parallel β-strands and one helix, located on top of the extended open sheet motif, is found [Figs. 9(f) and 11(a)]. The cluster is constituted by proteins binding either ATP, FAD, or NAD(P)(H). The central β-strand is oriented toward the connection between the adenine and ribose moiety and the α-helix points to the phosphate groups in the cofactors. The enzyme classification code suggests that more than 65% of the examples in this cluster are oxidoreductases.[83] Accordingly, it is not surprising that the extracted folding motif is part of the well-known Rossmann fold [Fig. 11(a)].

The second largest cluster 2 with 61 entries is comprised by proteins hosting ADP, ATP, and COA. It is structurally related to the motif found in cluster 1 and also shows a central five-stranded sheet [Fig. 11(a)]. This sheet is connected via helical elements above and below the sheet. Also here, one helix is oriented toward the diphosphate group of the cofactors. Mapping the Cavbase descriptors onto this motif shows five donor pseudocenters addressing the diphosphate group of the ligands. Of these, four are formed by backbone NH groups and one by the ammonium group of a conserved lysine. The EC numbers indicate that the majority of cofactors in this cluster originate from transferases and hydrolases. To estimate the affinity contribution of the different parts of the cofactors, an evaluation based on Drugscore[21] has

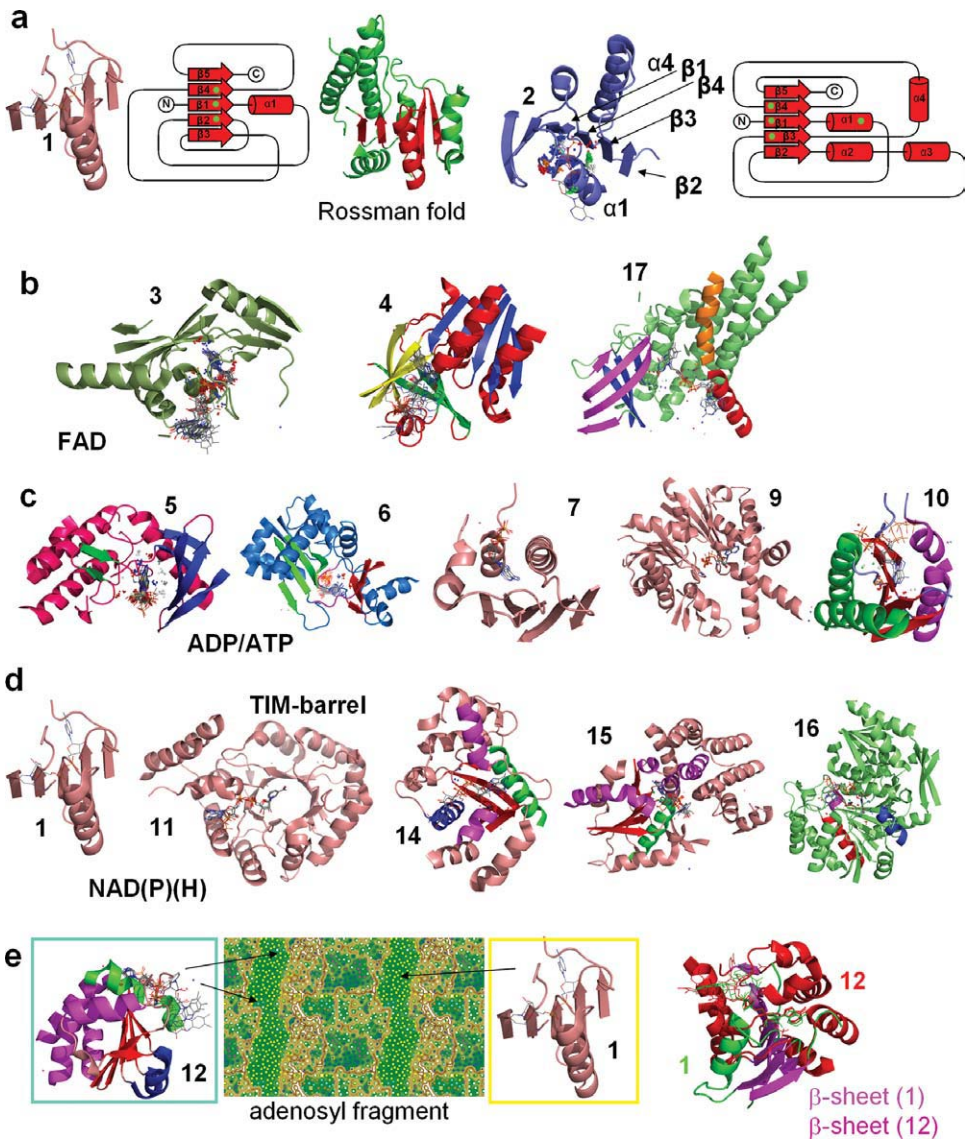

Figure 11

Clusters found in fold space: (**a**) In cavity and fold space two major clusters are indicated. Cluster 1 (left, beige) shares a local fold motif of a five-stranded parallel β-sheet with an adjacent similarly oriented helix covering the sheet in common. This motif is a substructure of the well-known Rossmann fold found in oxidoreductases (center, green, substructure motif of cluster 1 highlighted in red). In cluster 2 (center, blue) also a pattern with a parallel β-sheet is found, here encompassed by four helices. The topology clusters 1 and 2 differ by the connections of sheets and helices. (**b**) FAD is recognized by proteins in clusters 3, 4, and 17. They show distinct fold motifs and interestingly they also recognize different portions of the bound cofactors (cluster 3: adenosine moiety, cluster 4: isoalloxazine portion, and cluster 17: entire cofactor). (**c**) ADP and ATP are recognized by several local fold motifs (clusters 5, 6, 7, 9, and 10); cluster 5 represents the typical kinase fold. (**d**) Apart from cluster 1 (left) comprising a substructure of the Rossman fold, NAD(P)(H) is recognized by a TIM-barrel fold and three other folds exhibiting a mixed pattern of β-strands and α-helices. (**e**) The proteins in cluster 12 are capable to recognize different cofactors, however always through their adenosyl diphosphate unit. In these protein complexes, the adenosyl fragment adopts a conformation that is similar to the one bound by the five-stranded β-sheet-helix motif observed in cluster 1. An ESOM map trained with the torsion angles found for the adenosyl fragment indicates similar conformations for the cofactors bound to cluster 1 (best-matches in yellow) and cluster 12 (best-matches in cyan). Superposition of the local fold motifs from cluster 1 and 12 using the coordinates of the atoms of the adenosyl fragments reveals common orientation of a five-stranded β-sheet (light and dark magenta), but the helical arrangement (green and red) deviates strongly among the two fold motifs.

been performed. Drugscore is a knowledge-based scoring function and sums over the estimated affinity contributions between ligand atoms and adjacent protein atoms. Applying Drugscore separately for the three cofactors ADP, ATP, and CoA in this cluster, the portion comprising the diphosphate group is scored best. This suggests that the diphosphate group contributes most to the binding affinity in this cluster.

The clusters 3 (15 entries), 4 (8 entries), and 17 (3 entries) are all formed by proteins hosting FAD. Most examples belong to the class of oxidoreductases; in the third cluster also an isomerase is found. The folding pattern strongly deviates between the clusters and cofactor binding is achieved by functional groups located on different secondary structure elements [Figs. 10 and 11(b)]. Interestingly enough, in cluster 3 the adenosine phosphate moiety of FAD is addressed, whereas in cluster 4 the isoalloxazine moiety is recognized. In the third cluster 17, the entire cofactor is addressed and all three FADs adopt the same overall conformation. The affinity estimation with Drugscore suggests stronger binding contributions through the adenosine phosphate fragment in cluster 3. In case of cluster 4, the binding contribution originating from the isoalloxazine portion seems not significantly larger than for the remaining parts of the cofactor.

Clusters 5 (14 entries), 6 (12 entries), 7 (4 entries), 9 (7 entries), and 10 (5 entries) are all populated by proteins hosting either ATP or ADP [Figs. 10 and 11(c)]. The fold motifs in the first three clusters share antiparallel β-sheets in common; in cluster 7 one of the bordering β-strands of the β-sheet is attached with parallel orientation. The β-sheet in cluster 10 is exclusively oriented in parallel fashion. Cluster 5 comprises only transferases, mainly composed of kinases and one aminoglycoside phosphotransferase (1j7l) with striking homology to eukaryotic protein kinases. Here, the cofactor molecules superimpose well and exhibit virtually the same overall conformation with some scatter about the terminal phosphate group. Cluster 6 comprises different types of enzymes, for example, 5/6-kinases, but also synthases and carboxylases from the class of ligases are found. The cofactor molecules are recognized between two antiparallel β-sheets formed by four β-strands each. According to the Cavbase model, the ligands in this cluster are mainly addressed through the adenine and phosphate moieties. In cluster 7, two His kinases (1i58 and 2c2a), the chaperone HSP90 (1byq), and a pyruvate dehydrogenase kinase (2zkj) are found. The cofactor molecules are oriented differently to the adjacent β-sheet, and on the opposing side three α-helices close up the recognition site. In the Cavbase model as consensus binding area, the adenine moiety and the diphosphate group are addressed. The adenosine is embedded in a helical environment (α7–α11) and the diphosphate group is next to the edge of the neighboring β-strands. The fold motifs in clusters 9 (actin and actin homolog proteins, two chaperones HSP70 and HSC70) and 10 (ligases) are rather different. Although in cluster 9 all cofactor molecules adopt a rather similar conformation, in cluster 10 only the adenosyl fragments are well aligned.

The clusters 11 (4 entries), 14 (4 entries), 15 (4 entries), and 16 (3 entries) comprise proteins that recognize NAD(P)(H) as cofactor [Figs. 10 and 11(d)]. They show deviating fold topology next to the cofactor-binding pocket. In cluster 1, a substructure of the Rossmann fold has already been detected as binding-competent motif. Cluster 11 contains examples from the aldo-keto reductase superfamily but also the beta-subunit (kcnab2) from the human potassium channel Kv. They show TIM-barrel fold formed by eight α-helices and eight β-sheets. The cofactor molecules adopt extended conformation with the nicotine amide moiety next to the center of the barrel. The adenine terminus orients toward the helical rim of the barrel. Clusters 14 (adenylyltransferases) and 15 [oxidoreductases, 3-glycerol (1jq5) and 1,3-propanediol dehydrogenase (1o2d), and the lyases dehydroquinate (1sg6) and 2-deoxy-scyllo-inosose synthase (2gru)] exhibit parallel oriented β-sheets of five and six strands. The cofactor recognition site is complemented by helical segments. In cluster 14, they orient toward the adenine moiety (magenta, α5 and α8), run parallel to the latter group (green α1), or point to the nicotinamide (blue α4). In cluster 15, one helix orients toward the diphosphate group (green, α5) and four helices enclose the binding site of the adenine portion (magenta, α2, α3, α6, and α7). In cluster 16, two protein chains contribute to the cofactor-binding site. A turn motif is found next to the diphosphate group (brown) and the adenine fragment is flanked by two helices (red and magenta, α8 and α9). Another helix is oriented toward the nicotinamide portion (blue, α3).

Finally, cluster 12 (7 entries) is interesting as it recognizes NADP(H) in two examples of the studied data set, NAD(H) in one, ADP in one, and FAD in three cases. It comprises a lyase, an electron transfer protein, two oxidoreductases, and two hydrolases. The versatile recognition features of the motif in this cluster can be explained by the fact that only the adenosyl fragment of the different cofactors is bound with common orientation by the fold motif of this cluster. A six-stranded β-sheet (red) orients toward the adenine and ribose units. Two helices (green, α2 and α7) bind the diphosphate group. They are complemented on this side of the β-sheet by three additional helices (magenta, α1, α4, and α8) and on the opposite side of the sheet two further helices are found (blue, α6 and α7).

DISCUSSION

Torsion angle space

The analysis of conformational properties of the different cofactors clearly indicates several conformational families. The detailed evaluation of the six genuine torsion angles in 487 adenosine diphosphate moieties reveals torsional distributions in agreement with the results of previous investigations.[29,30] About the different single bonds mainly both *gauche* and the *anti-trans* conformations are populated. The ring pucker of the ribose ring

reveals two distinct conformers (C2′- and C3′-*endo*). The adenine portion adjusts preferentially with *anti* orientation, which projects the purine moiety optimally into protein environment. Nevertheless, some protein structures are known that exhibit reverse *syn* orientations of this ring system, for example, found in some of the CoA examples. Data analysis indicates that entries extracted from small-molecule crystal structures are distinct from those in protein complexes and fall in separate clusters. We evaluated the distance between the adenine moiety and the second P-atom of the pyrophosphate group [Fig. 2(d)], which adopts a shorter distance in the CSD entries (Fig. 3, left). This finding has been related to an "open" and "close" conformation by Saenger and is expressed by a distinct conformational distribution of τ_3 in small molecule and protein-bound entries.[30] Visual inspection of the former structures shows that the adenosine diphosphate moieties found in the small-molecule crystal structures are involved in the coordination of metal ions; in other cases, water molecules in the crystal packing mediate an interaction between adenosine and phosphate moiety leading to the observed back-folded and thus more compact geometries with shorter distances. Also, in the PDB data a bimodal distribution is observed, which decomposes, as the individual analyses show, into entries with large (8–10 Å, clusters 1 and 2, Figs. 2 and 3) and short distance (6–7.5 Å, clusters 3, 4, and 7). Interestingly, the expanded cofactors such as NAD(P)(H) or FAD are found more frequently with the larger distances, whereas the smaller ADP or ATP seem to populate more often with the smaller distances.

Cavity and fold space

To detect similarities among the proteins hosting one of the six cofactors considered in this study, one would at first evaluate sequence homology. To avoid, however, any at first glance trivial similarity among the studied proteins, we considered only examples with a sequence homology below 25%. Beyond this threshold, sequence data are no longer deemed as straightforward indicator for functional correlations between proteins. In consequence, we consulted approaches considering the spatial structure or architecture of the cofactor-binding sites. It is textbook knowledge that similarities in the fold of proteins are more pronouncedly conserved than sequences and correlate with function.

We performed analyses using Cavbase and DALI. The first method compares patterns of exposed physicochemical properties in binding pockets. It takes into account neither the architecture of the overall protein fold nor the contiguous spatial arrangement of secondary structure elements, which contribute to the construction of a binding site. DALI starts from a different point of view. This method compares proteins in terms of their fold geometry and because the spatial protein structure is determinant for its function, often similar folds parallel with similar functions. The Cavbase approach has been demonstrated to reliably classify proteins in terms of functional similarity in several studies. This fact might be surprising as the method only "looks at the first layer" of amino acids comprising a binding pocket and not at adjacent fold motifs. This study reveals, interestingly enough, a similar clustering either by Cavbase or by DALI of the 367 proteins, hosting one of the six considered cofactors. The degree of similarity is quantitatively not exactly the same among the different entries, but the detected cluster structure agrees well. One explanation for this observation is the fact that the clustering in cavity space works remarkably well if the analysis focuses only on the pseudocenters as descriptors that originate from functional groups residing on the backbones. These parts of the proteins correlate indirectly with the information evaluated by DALI. If the corresponding information is evaluated based on the pseudocenters assigned to the side-chain atoms, a more blurred clustering is detected in cavity space (Supporting Information Fig. S7).

The fold motifs, as identified by DALI next to the cofactor-binding sites, are substructures of the overall fold of the individual proteins. The motifs found in the two largest clusters 1 and 2 are reminiscent of the Rossmann fold. Once the local fold motifs have been extracted, we checked whether among the amino acids contributing to these common motifs, a higher sequence homology is given than the 25% applied as initial threshold. We could not detect, however, any significantly enhanced sequence similarity across these motifs. Possibly, this relates to our findings in cavity space that pseudocenters attributed to the side-chain atoms reveal a much less clear-cut clustering of the data.

The analysis of the binding pockets, particularly focusing on the subcavities hosting only fragments of the cofactors, clearly reveals that common conformational properties in some parts of the ligands also correlate with common structural patterns in the respective subpockets. They recognize conformationally alike ligand portions. The use of a method to predict binding affinity (here the scoring function Drugscore) suggests that in the latter cases also an enhanced contribution to the binding affinity can be expected from this commonly oriented and similarly recognized portion of the ligands.

Common features found by mapping across all spaces

In the following, we will select some illustrative examples to show how the properties observed between the different spaces mutually correlate and allow for some instructive conclusions. In Figure 2(a,b), the conformational analysis of the adenosyl fragment has been presented. In Figure 2(c), the best-matches on this ESOM map are color-coded with respect to the clustering found

in cavity space (Fig. 6). A fair number of entries remain unassigned (white); these correspond to either CSD entries or examples not considered in the Cavbase study due to incomplete embedding in a binding pocket. Examples from the large Cavbase cluster 1 [yellow, Fig. 2(c)] populate overwhelmingly in the central ESOM cluster [Fig. 2(a), cluster named 2 on the ESOM]. This correlation indicated that adenosyl fragments, mainly originating from NAD(P) (H) and FAD, share a common conformation and all bind to a fold motif constructed from a five-stranded parallel β-sheet with one parallel helix on top of the sheet. A second rather homogeneously populated cluster on the ESOM map is indicated by the blue-colored best-matches [Fig. 2(c)]. They correspond to entries falling into the second largest cluster in cavity space (Fig. 6). Here, the adenosyl fragment has been extracted from either ADP or ATP as cofactor, and the conformational analysis shows that the adenosyl fragments are all superimposed by their diphosphate groups [Fig. 2(c)]. We, therefore, inspected the ADP cofactor molecules closer. Their conformational analysis has been presented in Supporting Information Figure S2 and shows a large cluster on the ESOM with ligands adopting overall a very similar conformation. Mapping the Cavbase classification onto the best-matches on this ESOM map (Supporting Information Fig. S13) reveals that many entries assigned to the second Cavbase cluster actually populate in this valley on the ESOM map.

The third largest cluster in cavity space (Fig. 6) is exclusively populated by FAD ligands. They are indicated on the ESOM conformational map, computed for the adenosyl fragment, by the magenta-colored best-matches [Fig. 2(c)]. The corresponding ligands are superimposed through a common orientation of their adenosyl diphosphate portions. We, therefore, consulted the ESOM map for FAD again (Supporting Information Fig. S4). This map has highlighted one major cluster (Supporting Information Fig. S14), which, interestingly enough, hosts mostly cofactor molecules populating the Cavbase or DALI clusters 1 and 17 (Fig. 11). The FAD molecules are conformationally rather similar in this cluster. The FAD cofactors found in clusters 3 and 4 of cavity and fold space bind to distinct fold pattern. In cluster 3, the ligands are mainly recognized through their adenosyl fragments. On the ESOM, they fall (magenta-colored best-matches) in a similar region of the map but do not form a homogeneous cluster (Supporting Information Fig. S14). Very similar findings can be attributed to cluster 4 found in cavity and fold space. Here, the fold motif recognizes FAD via its isoalloxazine portion. On the ESOM, the entries of this cluster fall next to each other in a roughly indicated common cluster region (Supporting Information Fig. S14). To analyze these differences in detail, we trained an ESOM based on the FAD ligands only (Supporting Information Fig. S15) and considered solely the torsion angles of the adenosyl fragment [torsion angle definition, see Fig. 2(d)]. A large cluster is obtained, predominantly populated by the entries of the first cavity or fold cluster (Fig. 6). However, a second cluster is revealed with nine entries (magenta-colored best-matches) from the third and two entries (cyan-colored best-matches) from the fourth cavity cluster. Plotting the similarity matrix for the latter 11 entries, either according to Cavbase or DALI, reveals a clear separation of these examples. The differences in the protein environment become even more apparent when the protein structures are mutually superimposed taking the atoms of the adenosyl fragment as common bases for the mutual alignment (Supporting Information Fig. S16).

Cluster 12 in the cavity and fold space hosts either FAD, NAD(P)(H), or ADP as ligands. This versatile recognition is achieved by addressing a common binding motif of the adenosyl substructure. Considering the ESOM map of this fragment [Fig. 11(e)] shows one extended cluster with either entries corresponding to the large Cavbase cluster 1 (best-matches colored in yellow) and entries from cluster 12 in the cavity space (best-matches in cyan). Superimposing the fold motifs from clusters 1 and 12 based on the coordinates of the atoms of the adenosyl fragment shows common arrangement of a parallel β-sheet but distinct orientations of the surrounding helices. This explains why clusters 1 and 12, either in cavity or fold space, do not experience pronounced similarity.

In our study, FAD is found to be bound to entries from five different clusters (1, 3, 4, 12, and 17). Dym and Eisenberg[32] identified four deviating fold motifs for FAD-containing proteins (Rossmann-like in GR: glutathione reductase; PO: pyruvate oxidase; cylindric β-domain in FR: ferredoxin reductase; and PCMH: P-cresol methylhydroxylase). Examples of these four classes are found in our clusters 1 (1b37), 12 (1o97), 4 (1fdr, 1cqx), and 3 (1f0x).

Relationships among entries found in the unassigned region

The above-described examples give an idea about inherent correlations that become obvious through the comparative analysis of the different spaces. Finally, we want to focus our similarity analysis in cavity and fold space on the 97 structures that remained unassigned to any of the 17 clusters (Fig. 6, magenta box). This area is mainly assigned to singletons. To detect putative pairwise relationships in this area, we trained an ESOM map with these 97 entries based on the torsion angles of the adenosyl fragment. The obtained map shows some clustering. In the following, the coordinates of the atoms of the commonly matched adenosyl fragments were used to align the hosting proteins and to search for fold and cavity similarities. This way, 10 pairs could be identified that show similarity in either cavity or fold space (Fig. 12). Clusters III and IV do not indicate high similarity, and

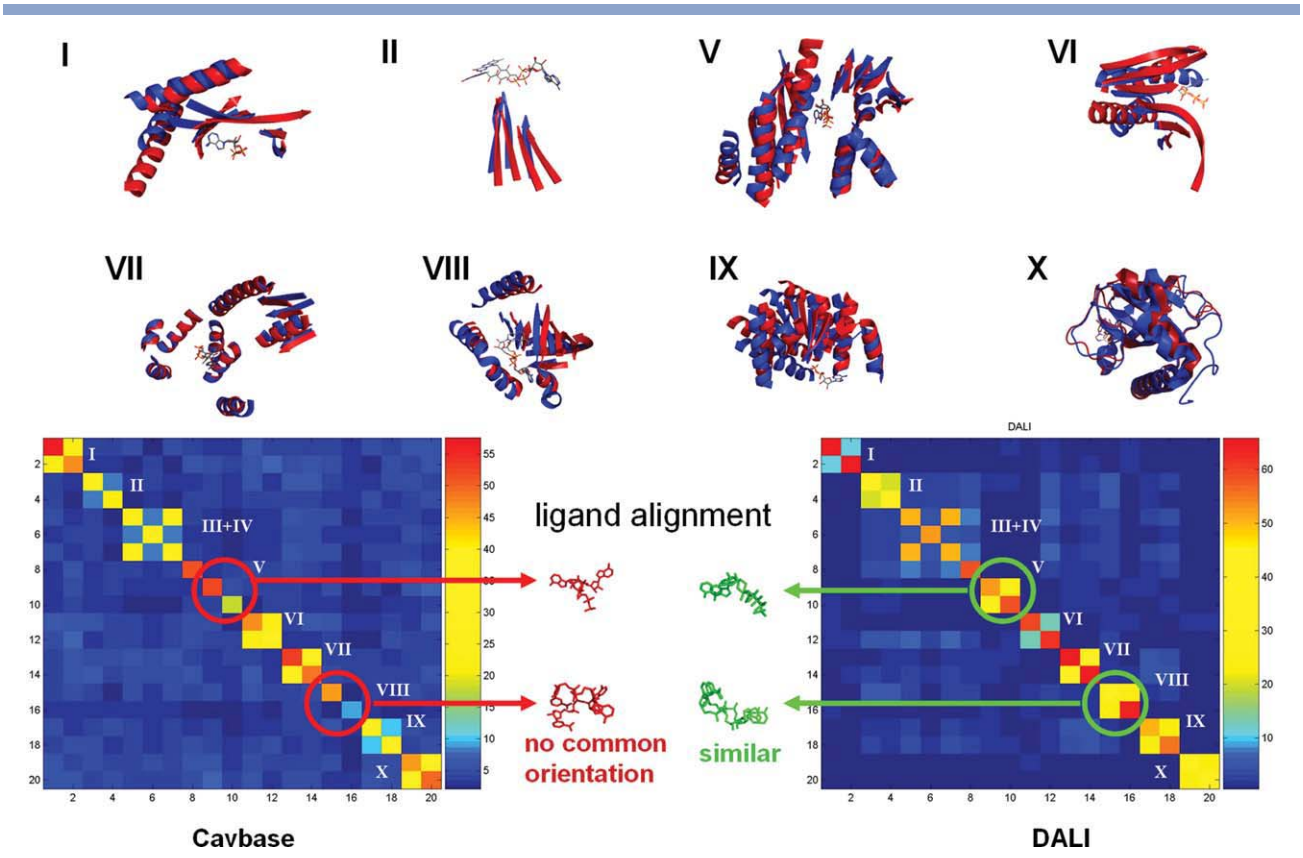

Figure 12

Using the results of a cluster analysis in torsion angle space for adenosyl fragments taken from entries in the unclassified region in cavity or fold space (Fig. 6, magenta box) suggests, nevertheless, some relationships between the hosting proteins. The results of this analysis retrieve pairs for related proteins that exhibit similar local fold motifs (clusters I, II, V, VI, VII, VIII, IX, and X; on the top, red and blue structures superimposed). The cavity comparison based on Cavbase descriptors (bottom, left) does not detect reasonable similarity among the pairs of proteins in clusters V and VIII (red cycles), whereas DALI (bottom, right) detects a significant similarity (green cycles). Superimposing the ligand coordinates of the bound NAD(P)(H)s achieves—with the transformation suggested by DALI—a reasonable superposition (green ligands), whereas the transformation suggested by Cavbase fails to produce a meaningful ligand alignment (red ligands). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

the corresponding proteins show β-sheets that deviate obviously too strongly to be detected by DALI as significantly similar. Clusters V and VIII are indicated by DALI as similar in fold space; however, Cavbase does not return this similarity. It is remarkable that the superposition of the bound NAD(P)(H) ligands (green) achieves reasonable matching when the transformation is computed using the results of DALI, whereas the superposition based on the cavity alignment cannot detect a comparable orientation of the bound cofactors (red). This latter example demonstrates the strength of an approach combining the information from torsion angle, cavity, and fold space to elucidate similarities among proteins with remote sequence similarity.

CONCLUSION

Proteins that bind similar ligands or ligand portions must exhibit common structural motifs or patterns of physicochemical properties in their recognition pockets. Such properties make them competent to interact with the same ligands and can be responsible for undesired side effects of drug molecules. This study has focused on cofactor-binding sites, as a sufficient body of crystal structures is available to perform statistical analyses. Only proteins with sequence homology below a threshold of 25% have been considered to exclude trivial correlations.

The torsion angle analysis reveals distinct clustering concerning conformational preference about the *N*-glycosidic bond, the ribose ring pucker, and the various *gauche*- and *trans*-orientations about different single bonds. Overall, a clustering in several conformation families is observed.

Remarkably, the clustering in cavity and fold space returns a very similar clustering pattern albeit, at some points with somewhat different ranking. Some of the clusters are uniquely populated by one cofactor chemotype, whereas others are competent to recognize several distinct cofactor types. The analysis clearly shows that in

some of the latter (mixed) clusters only parts of the cofactors are recognized showing common orientation and these parts give rise to the obtained (mixed) clusters with common patterns.

In fold space, we extracted commonly shared fold motifs from the overall structures. As most populated fold motif, a parallel oriented β-sheet structure, flanked by helices adjacent to the sheet, is found. This pattern reminds of the Rossmann fold, observed in many oxidoreductases, which is well known to bind NAD(P)(H) as cofactor.[83] Nevertheless, also completely different fold motifs were detected, for example, the TIM barrel or antiparallel β-sheets that also bind NAD(P)(H). In total, we could detect five different fold motifs (clusters 1, 11, 14, 15, and 16) to interact with the latter cofactor type (Fig. 11). With respect to FAD, the analysis also returns five different fold motifs (clusters 1, 3, 4, 12, and 17), which are competent to recognize this cofactor. Most widely recognized is ADP/ATP as we found examples for binding proteins in clusters 1, 2, 5, 6, 7, 9, and 10. Also, clusters containing proteins competent to recognize different cofactors are discovered, for example, cluster 12. Interestingly enough, here only the part that is shared in common by all cofactors binds to the fold motif, and this portion adopts the same conformation in all cofactors.

Particularly, the mapping of the clustering pattern in one space onto the analysis in the other spaces (e.g., torsion angle space vs. cavity and fold space) detects some quite remarkable features and helps to discover even remote relationships. This approach appeared most powerful to find correlations among proteins classified at first as singletons. Here, some pairs of proteins could be detected that share local fold geometries in common.

The Cavbase approach returns surprisingly well a very similar relationship compared to the fold space analysis using DALI. Supposedly, this finding results from the many Cavbase descriptors that are contributed by the functional groups residing on the backbone. This obviously gives a sufficient description of the secondary structure elements important for the local folding of a protein. The Cavbase approach appears powerful to detect similarities across ligand-binding sites and to cluster them successfully. However, one major problem, particularly in predicting cross reactivity at remote proteins to estimate possible side effects of drug molecules, is first conformational flexibility of the ligands, which leads to the facile adoption of a different shape, allowing them to bind to pockets of deviating architecture. Second, this observation seems to be even more puzzling, commonly shared recognition features in binding pockets need not extend over the entire ligands; sometimes only portions of the ligands are detected by common binding pocket patterns, which then induce a similar bound conformation.

MATERIALS AND METHODS

Datasets

Using Relibase,[84] all protein complexes containing one of the cofactors ADP, ATP, NAD(P)(H), FAD, and acetyl CoA (Table I) have been extracted. The considered crystal structures were required to have a resolution of at least 2.3 Å; the ligands had to be fully populated and structures showing *B*-factors > 50 Å^2 were discarded. To avoid redundancies in sequence space, all protein chains contributing to the cofactor-binding site (<5 Å distance from one atom of the cofactor) were collected according to the processed data set based on the PDB select list.[85] Only structures with less than 25% homology were considered. Among the entries exhibiting larger mutual sequence homology, one entry has been selected as representative. For this selection, the entry with highest crystallographic resolution and lowest values of the *B*-factors was considered. After application of these selection criteria, 459 individual entries (extracted from 452 crystal structures) remained for the subsequent analysis (Supporting Information Table S1).

For the purpose of comparison, also data from small-molecule crystal structures (CSD)[86] have been retrieved. As query fragment, an adenosine diphosphate moiety has been searched and 28 CSD entries have been added to the torsion angle analysis. Therefore, the analysis in torsion angle space comprised a dataset of 487 entries (Supporting Information Table S1).

For the binding pocket analysis, all ligands considered in the torsion angle analysis were required to be fully embedded in a protein-binding pocket. Within a 4 Å sphere, at least one atom of the cofactor had to coincide with a grid point defining the surface of the cavity as extracted by the program LIGSITE.[60] This criterion was applied on purpose because in a subsequent analysis in binding pocket space a splitting of the cofactors in individual building blocks (Table I) was intended. After applying these criteria, a reduced dataset of 367 entries was considered in the cavity- and fold-space evaluation (Supporting Information Table S3).

Torsion angle analysis

For the ligands considered in the torsion-angle analysis, the individual torsion angle values have been extracted and tabulated using Relibase[84] and Conquest.[86]

ESOM analyses

Clustering of the torsion angles has been performed using ESOMs[38] as previously described by us.[37] For each ligand type, a toroidal net with 50×80 neurons has been applied. A total of 400 epochs have been used for each training run. The detailed parameter settings of the ESOM analysis are given in Supporting Information

Table S4. The results of the ESOM analyses are displayed graphically using a kind of "3D landscape of valleys and mountains" based on the values of the corresponding U-matrix. Onto this map, the individual entries can be projected as best-matches[39] and, using color-coding for the best-matches, additional properties, such as the chemotype of the cofactor, a particular torsion angle value, or the membership to a cluster found in cavity space, can be indicated. This helps to detect easily where particular entries with a given property are clustered on a map.

Pareto density estimation

Pareto density estimation has been used as described by Ultsch[87] to evaluate the torsion angle scatter in terms of a continuous density distribution. This estimation allows assigning a continuous function to the scatter of individual observations (total number of observations d). The data distribution $P(x)$ at a given point x is defined by eq. 1, with NN being the number of nearest neighbors with a predefined radius r. This so-called Pareto radius is set, according to Ultsch,[88] in a way that the median of NN(x,r) comprises 20% of all data points. In the Supporting Information, the computed Pareto density estimations for a trimodal torsion angle distribution are given based on different bin sizes to represent the torsional data (Supporting Information Fig. S17). With this technique, a continuous data distribution is generated, which is quite independent of the predefined bin width.

Cavbase similarity analysis and clustering of Cavbase descriptors

The mutual comparison of binding pockets is performed using Cavbase.[18] Common spatial arrangements of the cavity-assigned pseudocenters are found using the Bron-Kerbosch clique detection algorithm.[89] Subsequently, the cavities are aligned based on the detected cliques. For each alignment, a surface overlap score is calculated based on the degree to which surface patches, assigned to the same pseudocenter type in the two cavities, overlap with one another. The cavity surface patches are based on the cavity flanking grid points determined by LIGSITE.[60] Cavbase evaluates the surface overlap score for the 100 largest cliques and the alignment leading to the best score is stored.

For the following comparative analysis, each cavity in the dataset is compared one-by-one to all other cavities. Alternative scoring schemes have been suggested to rank and subsequently cluster the best scored superpositions.[18] In this study, we applied the scoring scheme R_3, as previously described by Kuhn *et al.*[90] The resulting scores are stored in a similarity matrix that serves as input for various clustering algorithms. Clustering has been performed by agglomerative hierarchical clustering using Ward's criterion.[91] This protocol has been applied either to all 367 surround pockets of the data set simultaneously, to the surround cavities of the six cofactors separately, or to the different subcavities obtained after splitting the cofactors into fragments. In all cases, a reasonable estimate of the number of obtained clusters has been achieved by closely inspecting the corresponding dendrograms (cf. Supporting Information Figs. S8 and S9) of the agglomerative hierarchical clustering. The similarity matrices displayed graphically in the various figures (e.g., Fig. 5 or 7) were composed and sorted according to these clustering results, and a color-coding corresponding to the percentage of achieved similarity has been assigned. The graphical images were produced using the program Matlab.[92]

Classification of the folding topology by TOPS

Using the program TOPS, a protein topology cartoon has been produced for each fold cluster taking one representative cluster member as a reference.[81,82] The program generates a schematic representation of the input protein structure as a string of secondary structure elements that preserves connectivity and directionality (β-strands as arrows and α-helices as cylinders).

Classification according to EC classification code

The EC classification code has been extracted for the selected examples from Relibase.[84]

ACKNOWLEDGMENTS

The help of Prof. Ultsch (Univ. Marburg), particularly at the beginning of the project in providing B.S. an introduction to the use of ESOM maps, is gratefully acknowledged. The authors thank Florian Meyer (Univ. Marburg) who supported B.S. in many questions concerning clustering and data analysis. The support of Gerd Neudert and Sven Siebler (Univ. Marburg), who made their computer tools, particularly Drugscore and Cavbase routines, available to B.S., is kindly acknowledged. G.K. thanks Michael Betz (Univ. Marburg) for his help in retrieving essential information from the thousands of remaining files of B.S. after his sudden death. Due to an accident in 1999, B.S. was paralyzed from the sixth cervical vertebra onwards and wheelchair-bound. He did not give up and several years later he found the enormous power to embark into data analysis of protein-ligand complexes, a field completely new to him. We admired his incredible wish to realize his thesis despite many nasty health problems. All of a sudden B.S. passed away. We are grateful to all who supported him in this time, particularly Clemens

Schwan (Univ. Marburg), the German Rehabilitation Fond and Herrmann-Leuschner Stiftung.

REFERENCES

1. Kortagere S, Krasowski MD, Ekins S. The importance of discerning shape in molecular pharmacology. Trends Pharmacol Sci 2009;30: 138–147.
2. Kahraman A, Morris RJ, Laskowski RA, Thornton JM. Shape variation in protein binding pockets and their ligands. J Mol Biol 2007;368:283–301.
3. Amyes TL, Richard JP. Rational design of transition-state analogues as potent enzyme inhibitors with therapeutic applications. ACS Chem Biol 2007;20:711–714.
4. Mestres J, Gregori-Puigjane E, Valverdebc S, Sole RV. The topology of drug–target interaction networks: implicit dependence on drug properties and target families. Mol BioSyst 2009;5:1051–1057.
5. Bode W, Turk D, Stürzebecher J. Geometry of binding of the benzamidine- and arginine-based inhibitors N^{α}-(2-naphthyl-sulphonyl-glycyl)-dl-*p*-amidinophenylalanyl-piperidine (NAPAP) and (2*R*,4*R*)-4-methyl-1-[N^{α}-(3-methyl-1,2,3,4-tetrahydro-8-quinolinesulphonyl)-l-arginyl]-2-piperidine carboxylic acid (MQPA) to human α-thrombin. X-ray crystallographic determination of the NAPAP-trypsin complex and modeling of NAPAP-thrombin and MQPA-thrombin. Eur J Biochem 1990;193:175–182.
6. Banner DW, Hadváry P. Crystallographic analysis at 3.0-A resolution of the binding to human thrombin of four active site-directed inhibitors. J Biol Chem 1991;266:20085–20093.
7. Weber A, Casini A, Heine A, Kuhn D, Supuran CT, Scozzafava A, Klebe G. Unexpected nanomolar inhibition of carbonic anhydrase by COX-2-selective celecoxib: new pharmacological opportunities due to related binding site recognition. J Med Chem 2004;47:550–557.
8. Kawamori T, Rao CV, Seibert K, Reddy BS. Chemopreventive activity of celecoxib, a specific cyclooxygenase-2 inhibitor, against colon carcinogenesis. Cancer Res 1998;58:409–412.
9. Oshima M, Murai N, Kargman S, Arguello M, Luk P, Kwong E, Taketo MM, Evans JF. Chemoprevention of intestinal polyposis in the Apcdelta716 mouse by rofecoxib, a specific cyclooxygenase-2 inhibitor. Cancer Res 2001;61: 1733–1740.
10. Waskewich C, Blumenthal RD, Li H, Stein R, Goldenberg DM, Burton J. Celecoxib exhibits the greatest potency amongst cyclooxygenase (COX) inhibitors for growth inhibition of COX-2-negative hematopoietic and epithelial cell lines. Cancer Res 2002;62:2029–2033.
11. Moreira L, Castells A. Cyclooxygenase as a target for colorectal cancer chemoprevention. Curr Drug Targets 2010;11(12).
12. Cowan-Jacob SW, Fendrich G, Floersheimer F, Furet P, Liebetanz J, Rummel G, Rheinberger P, Centeleghe M, Fabbro D, Manley PW. Structural biology contributions to the discovery of drugs to treat chronic myelogenous leukaemia. Acta Crystallogr D Biol Crystallogr 2007;63:80–93.
13. Winger JA, Hantschel O, Superti-Furga G, Kuriyan J. The structure of the leukemia drug imatinib bound to human quinone reductase 2 (NQO2). BMC Struct Biol 2009;9:7–7.
14. Istvan ES, Deisenhofer J. Structural mechanism for statin inhibition of HMG-CoA reductase. Science 2001;292:1160–1164.
15. Kallen J, Welzenbach K, Ramage P, Geyl D, Kriwacki R, Legge G, Cottens S, Weitz-Schmidt G, Hommel U. Structural basis for LFA-1 inhibition upon lovastatin binding to the CD11a I-domain. J Mol Biol 1999;292:1–9.
16. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
17. Taylor R. Life science applications of the Cambridge Structural Database. Acta Cryst D 2002;58:879–888.
18. Schmitt S, Kuhn D, Klebe G. A new method to detect related function among proteins independent of sequence and fold homology. J Mol Biol 2002;323:387–406.
19. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233:123–138.
20. Webb EC. Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes. San Diego: Academic Press; 1992.
21. Gohlke H, Hendlich M, Klebe G. Knowledge-based scoring function to predict protein-ligand interactions. J Mol Biol 2000;295: 337–356.
22. Boström J. Reproducing the conformations of protein-bound ligands: a critical evaluation of several popular conformational searching tools. J Comput-Aided Mol Des 2001;15:1137–1152.
23. Nicklaus MC, Wang S, Driscol JS, Milne GWA. Conformational changes of small molecules binding to proteins. Bioorg Med Chem 1995;3:411–428.
24. Perola E, Charifson PS. Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand recognition upon binding. J Med Chem 2004;47:2499–2510.
25. Klebe G, Mietzner T. A fast and efficient method to generate biologically relevant conformations. J Comput-Aided Mol Des 1994;8:583–606.
26. Hao M-H, Haq O, Muegge I. Torsion angle preference and energetics of small molecule ligands bound to proteins. J Chem Inf Model 2007;47:2242–2252.
27. Bramfeld KA, Kuhn B, Reuter DC, Stahl M. Small molecule conformational preferences derived from crystal structure data. A medicinal chemistry focused analysis. J Chem Inf Model 2008;48:1–24.
28. Butler KT, Luque FJ, Barril X. Towards accurate energy predictions of the bioactive conformation of drugs. J Comput Chem 2008;30:601–610.
29. Moodie SL, Thornton JM. A study into the effect of protein binding on nucleotide conformation. Nucleic Acids Res 1993;26:1369–1380.
30. Klebe G. Structure correlation and ligand/receptor interactions. In: Bürgi HB, Dunitz JD, editors. Structure correlation. Weinheim: VCH; 1994. pp 543–603.
31. Carugo O, Argos P. NADP-dependent enzymes. Conserved stereochemistry of cofactor binding. Proteins 1997;28:10–28.
32. Dym O, Eisenberg D. Sequence-structure analysis of FAD-containing proteins. Protein Sci 2001;10:1712–1728.
33. Stockwell G, Thornton JM. Conformational diversity of ligands bound to proteins. J Mol Biol 2006;356:928–944.
34. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH—a hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.
35. Mitchell JBO. The relationship between the sequence identities of alpha helical proteins in the PDB and the molecular similarities of their ligands. J Chem Inf Comput Sci 2001;41:1617–1622.
36. Fisher RA. On the interpretation of χ^2 from contingency tables, and the calculation of P. J R Stat Soc 1922;85:87–94.
37. Koch O, Klebe G. Turns revisited: a uniform and comprehensive classification of normal, open and reverse turn families minimizing unassigned random chain portions. Proteins 2009;74:353–367.
38. Ultsch A. Maps for the visualization of high-dimensional data spaces. In: Proceedings of the 4th Workshop on Self-Organizing Maps, Kyushu, Japan, 2003. pp 225–230.
39. Kupas K, Ultsch A, Klebe G. Classification of substructures in protein binding cavities using wavelets. Proteins 2008;71:1288–1306.
40. Sundaralingam M. Structure and conformation of nucleosides and nucleotides and their analogs as determined by Xray diffraction. Ann N Y Acad Sci 1975;255:3–42.
41. Saenger W. Structure and function of nucleosides and nucleotides. Angew Chem Int Ed Engl 1973;12:591–601.

42. Reddy BS, Saenger W, Mühlegger K, Weimann G. Crystal and molecular structure of the lithium salt of nicotinamide adenine dinucleotide dihydrate (NAD^+, DPN^+, cozymase, codehydrase I). J Am Chem Soc 1981;103:907–914.
43. Laskowski RA, Luscombe NM, Swindells MB, Thornton JM. Protein clefts in molecular recognition and function. Protein Sci 1996;5:2438–2452.
44. Campbell SJ, Gold ND, Jackson RM, Westhead DR. Ligand binding: functional site location, similarity and docking. Curr Opin Struct Biol 2003;13:389–395.
45. Gold ND, Jackson RM. Fold independent structural comparisons of protein–ligand binding sites for exploring functional relationships. J Mol Biol 2006;355:1112–1124.
46. Rosen M, Lin SL, Wolfson H, Nussinov R. Molecular shape comparisons in searches for active sites and functional similarity. Protein Eng Des Select 1998;11:263–277.
47. Whisstock JC, Lesk AM. Prediction of protein function from protein sequence and structure. Quart Rev Biophys 2003;36:307–340.
48. Nicholls A, McGaughey GB, Sheridan RP, Good AC, Warren G, Mathieu M, Muchmore SW, Brown SP, Grant JA, Nevins N, Jain AN, Kelley B. Molecular shape and medicinal chemistry: a perspective. J Med Chem 2010;53:3862–3886.
49. Baroni M, Cruciani G, Sciabola S, Perruccio F, Mason JS. A common reference framework for analyzing/comparing protein and ligands. Fingerprints for ligands and proteins (FLAP): theory and application. J Chem Inf Model 2007;47:279–294.
50. Henrich S, Salo-Ahena OMH, Huanga B, Rippmann F, Cruciani G, Wade RC. Computational approaches to identifying and characterizing protein binding sites for ligand design. J Mol Recognit 2010;23:209–219.
51. Kinoshita K, Furui J, Nakamura H. Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. Protein Sci 2003;12:1589–1595.
52. Najmanovich R, Kurbatova N, Thornton JM. Detection of 3D atomic similarities and their use in the discrimination of small molecule protein-binding sites. Bioinformatics 2008;24:i105–i111.
53. Laskowski RA. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J Mol Graph 1995;13:323–330.
54. Park K, Kim D. Binding similarity network of ligand. Proteins 2008;71:960–971.
55. Jambon M, Imberty A, Deléage G, Geourjon C. A new bioinformatic approach to detect common 3D sites in protein structures. Proteins 2003;52:137–145.
56. Powers R, Copeland JC, Germer K, Mercier KA, Ramanathan V, Revesz P. Comparison of protein active site structures for functional annotation of protein and drug design. Proteins 2006;65:124–135.
57. Pickering SJ, Bulpitt AJ, Efford N, Gold ND, Westhead DR. AI-based algorithms for protein surface comparisons. Comput Chem 2001;26:79–84.
58. Schalon C, Surgand JS, Kellenberger E, Rognan D. A simple and fuzzy method to align and compare druggable ligand binding sites. Proteins 2008;71:1755–1778.
59. Shulman-Peleg A, Nussinov R, Wolfson HJ. Recognition of functional sites in protein structures. J Mol Biol 2004;339:607–633.
60. Hendlich M, Rippmann F, Barnickel G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. J Mol Graph Model 1997;389:359–363.
61. Porter CT, Bartlett GJ, Thornton JM. The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res 2004;32:D129–D133.
62. Binkowski TA, Adamian L, Liang J. Inferring functional relationships of proteins from local sequence and spatial surface patterns. J Mol Biol 2003;332:505–526.
63. Binkowski TA, Naghibzadeg S, Liang J. CASTp: computed atlas of surface topography of proteins. Nucleic Acids Res 2003;31:3352–3355.
64. Kleywegt GJ. Recognition of spatial motifs in protein structures. J Mol Biol 1999;285:1887–1897.
65. Barker JA, Thornton JM. An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. Bioinformatics 2003;19:1644–1649.
66. Besl PJ, McKay ND. A method for registration of 3-D shapes. IEEE Trans PAMI 1992;14:239–256.
67. Nussinov R, Wolfson HJ. Efficient detection of three-dimensional motifs in biological macromolecules by computer vision techniques. Proc Natl Acad Sci USA 1991;88:10495–10499.
68. Stark A, Russell RB. Annotation in three dimensions. PINTS: patterns in non-homologous tertiary structures. Nucleic Acids Res 2003;31:3341–3344.
69. Russell RB. Detection of protein three dimensional side-chain patterns: new examples of convergent evolution. J Mol Biol 1998;279:1211–1227.
70. Milik M, Szalma S, Olszewski KA. Common structural cliques: a tool for protein structure and function analysis. Protein Eng 2003;16:543–552.
71. Wallac AC, Borkakoti N, Thornton JM. TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases, application to enzyme active sites. Protein Sci 1997;6:2308–2323.
72. Brakoulias A, Jackson RM. Towards a structural classification of phosphate binding sites in protein-nucleotide complexes: an automated all-against-all structural comparison using geometric matching. Proteins 2004;56:250–260.
73. Spriggs RV, Artymiuk PJ, Willett P. Searching for patterns of amino acids in 3D protein structures. J Chem Inf Comput Sci 2003;3:412–421.
74. Artymiuk P, Poirrette AR, Grindley HM, Rice DW, Willett P. A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. J Mol Biol 1994;243:327–344.
75. Ausiello G, Via A, Helmer-Citterich M. Query3d: a new method for high-throughput analysis of functional residues in protein structures" BMC Bioinformatics 2005;6(Suppl 4):S5.
76. Wangikar PP, Tendulkar AV, Ramya S, Mali DN, Sarawagi S. Functional sites in protein families uncovered via a objective and automated graph theoretical approach. J Mol Biol 2003;326:955–978.
77. May ACW, Johnson MS. Protein structure comparisons using a combination of a genetic algorithm, dynamic programming and least-squares minimization. Protein Eng 1994;4:475–485.
78. Piorrette AR, Artymiuk PJ, Rice DW, Willett P. Comparison of protein surfaces using a genetic algorithm. J Comput-Aided Mol Des 1997;11:557–569.
79. Lehtonen J, Denessiouk K, May ACW, Johnson MS. Finding local structural similarities among families of unrelated protein structures: a generic non-linear alignment algorithm. Proteins 1999;34:341–355.
80. DeLano WL. The PyMOL molecular graphics system. San Carlos, CA: DeLano Scientific; 2002.
81. Gilbert D, Westhead D, Nagano N, Thornton JM. Motif-based searching in TOPS protein topology databases. Bioinformatics 1999;15:317–326.
82. Westhead DR, Hatton DC, Thornton JM. An atlas of protein topology cartoons available on the World-Wide Web. Trends Biochem Sci 1998;23:35–36.
83. Branden C, Tooze J. Introduction to protein structure, 2nd ed. New York: Garland; 1999.
84. Hendlich M, Bergner A, Günther J, Klebe G. Relibase: design and development of a database for comprehensive analysis of protein-ligand interactions. J Mol Biol 2003;326:607–620.
85. Hobohm U, Sander C. Enlarged representative set of protein structures. Protein Sci 1994;3:522–524.

86. Bruno IJ, Cole JC, Edgington PR, Kessler M, Macrae CF, McCabe P, Pearson J, Taylor R. New software for searching the Cambridge Structural Database and visualising crystal structures. Acta Cryst B 2002;58:389–397.
87. Ultsch A. Pareto density estimation: a density estimation for knowledge discovery. In: Baier D, Wernecke KD, editors. Innovations in classification, data science, and information systems—Proceedings of the 27th Annual Conference of the German Classification Society (GfKL) 2003. Berlin: Springer; 2003. pp 91–100.
88. Ultsch A. Eine Begründung der Pareto 80/20 Regel und Grenzwerte für die ABC Analyse, Technical Report No. 30, Department of Mathematics and Computer Science, University of Marburg, Germany, 2001.
89. Bron C, Kerbosch J. Algorithm 457. Finding all cliques of an undirected graph. Commun ACM 1973;16:575–577.
90. Kuhn D, Weskamp N, Schmidt S, Hüllermeier E, Klebe G. From the similarity analysis of protein cavities to the functional classification of protein families using Cavbase. J Mol Biol 2006;359:1023–1044.
91. Ward JH. Hierarchical grouping to optimize an objective function. J Am Stat Assoc 1963;58:236–244.
92. MATLAB technical documentation. Mathworks.com. Retrieved 2010-06-07.