Article No. mb981878 *J. Mol. Biol.* (1998) **280**, 535–550

JMB

Dictionary of Interfaces in Proteins (DIP). Data Bank of Complementary Molecular Surface Patches

Robert Preißner, Andrean Goede and Cornelius Frömmel*

Medical Faculty of the Humboldt University (Charité) Institute of Biochemistry Monbijoustr. 2A, Berlin D-10117, Germany

Molecular surface areas of proteins are responsible for selective binding of ligands and protein-protein recognition, and are considered the basis for specific interactions between different parts of a protein. This basic principle leads us to study the interfaces within proteins as a learning set for intermolecular recognition processes of ligands like substrates, coenzymes, etc., and for prediction of contacts occurring during protein folding and association. For this purpose, we defined interfaces as pairs of matching molecular surface patches between neighboring secondary structural elements. All such interfaces from known protein structures were collected in a comprehensive data bank of interfaces in proteins (DIP).

The up-to-date DIP contains interface files for 351 selected Brookhaven Protein Data Bank entries with a total of about 160,000 surface elements formed by 12,475 secondary structures. For special purposes, the inclusion of additional structures or selection of subgroups of proteins can be performed in an easy and straightforward manner. Atomic coordinates of the constituents of molecular surface patches are directly accessible as well as the corresponding contact distances from given atoms to their neighboring secondary structural elements.

As a rule, independent of the type of secondary structure, the molecular surface patches of the secondary structural elements can be described as quite flat bodies with a length to width to depth ratio of about 3:2:1 for patches consisting of more than ten atoms. The relative orientation between two docking patches is strongly restricted, due to the narrow distribution of the distances between their centers of mass and of the angles between their normal lines, respectively.

The existing retrieval system for the DIP allows selection (out of the set of molecular patches) according to different criteria, such as geometric features, atomic composition, type of secondary structure, contacts, etc. A fast, sequence-independent 3-D superposition procedure was developed for automatic searches for geometrically similar surface areas. Using this procedure, we found a large number of structurally similar interfaces of up to 30 atoms in completely unrelated protein structures.

Keywords: protein structure; secondary structure; interfaces; docking; folding

**Corresponding author*

Introduction

From a historical point of view, molecular biology encompasses a period of data collection and data banks are widespread therein. The Human Genome Project stimulates exponential growth of existing sequence data banks (Fasman *et al.*, 1994). The rate at which new protein 3D structures are being published is rapidly increasing too, but the difference between the number of known sequences and 3D structures in the Brookhaven Protein Data Bank (Bernstein *et al.*, 1977) is becoming larger (Orengo *et al.*, 1993). Due to the well-known fact that amino acid sequence homology at a given level leads to similar 3D structure

Abbreviations used: SSE, secondary structural element; MSP, molecular surface patch; ID, identifier; DIP, dictionary of interfaces in proteins; PDB, Brookhaven Protein Data Bank; vdW, van der Waals; H, helix; E, extended; C, coil; 3D, three-dimensional; rms, root-mean-square.

0022–2836/98/280535–16 $30.00/0

of proteins, several databases are interrelating the databases of sequences and structures: the homology derived structures of proteins (HSSP, Sander & Schneider, 1991), those of Pascarella & Argos (1992), SESAM (Huysmans *et al*., 1991) and the families of structurally similar proteins (FSSP, Holm *et al*., 1992). The search for homology in the sequence database is used to determine indications for function of proteins. The *a priori* knowledge of neighborhood and succession of amino acids in primary structures leads to less complex algorithms for sequence homology search, whereas the comparison of 3D structures at the atomic level first requires the assignment of equivalent atoms in the two sets (NP-hard problem; Kuhl *et al*., 1984), and then the superposition of these atoms. Taking this complexity into account, a number of different approaches have been applied (for a review, see Bures *et al*., 1994). The use of graph theory, differential geometry or atomic property information, as well as substructure searching or similarity screening, aims at a reduction of the number of dimensions and degrees of freedom. Up to now the algorithms most successfully applied for detection of similar spatial arrangements in proteins focus on the protein backbone (Vriend & Sander, 1991; Alexandrov *et al*., 1992; Lessel & Schomburg, 1994; Fischer *et al*., 1995; Alexandrov, 1996), which allows for classification of proteins (Alexandrov & Go, 1994). In a similar concept, any individual secondary structural element (SSE) is represented by a vector (Abagyan & Maiorov, 1992; Grindley *et al*., 1993; Mizuguchi & Go, 1995), through which database screening for similar arrangements of secondary structures becomes faster. Other approaches considering 3D similarities derive geometric descriptions like curvature, knobs and holes (Eisenhaber *et al*., 1995). However, none of these procedures takes into account the atomic position of amino acid side-chains. As a consequence, these methods suffer from a loss of atomic details responsible for molecular recognition during folding, binding, etc. The vector-representation of secondary structures (Alesker *et al*., 1996) works very fast for detection of super-secondary structural motifs and takes side-chains into account in a last refinement cycle, but nevertheless may be less successful in detecting geometric similarity in atomic detail without similarly arranged secondary structural elements. Due to a combinatorial explosion caused by the large number of atoms in proteins, methods developed to search for 3D similarity of low molecular mass substances often fail during estimation of equivalent atoms in proteins.

Molecular interaction is characterized by complementarity at the atomic level. This phenomenon is observed as physiochemical complementarity and as geometrical fitting of molecular surface areas. The latter is assumed to dominate the docking process (Connolly, 1986). The representation of molecular surfaces has to satisfy several requirements to be useful in docking prediction (Lawrence *et al*., 1987) and its calculation method has to be fast. The comparison of the general shape (e.g. concave, convex) should be possible as well as the comparison at the atomic level. Although analytical equations and fast algorithms for the calculation of surfaces have been derived (Connolly, 1983, 1985; Richmond, 1984; Eisenhaber & Argos, 1993; Eisenhaber *et al*., 1995), several methodological problems remain: choice of atom and probe radii, differences between surface calculations within proteins and at their exterior, errors during allocation of space among atoms (Gerstein *et al*., 1995; Goede *et al*., 1997).

Complementarity of paired molecular surfaces is relevant to ligand binding as well as to protein folding. The process of molecular recognition occurs through the association of complementary parts of molecules. The relationship of complementary surface areas is also relevant in folded proteins, between secondary structures and domains, between subunits (Argos, 1988), in dimeric structures (Jones & Thornton, 1995), in oligomeric proteins (Miller, 1989; Tsai *et al*., 1996), and in enzyme-inhibitor substrate (Janin & Chothia, 1990), or immunoglobulin-antigen complexes (Padlan, 1990). A similar classification of different protein-protein interactions was recently carried out by Jones & Thornton (1996). Recent studies of protein folding (e.g. see Peng *et al*., 1995) indicate that molten globules do have major secondary structures resembling native topology. These results suggest the existence of a intermediate state containing pre-formed secondary structural units that have to be closely packed during further folding. The concept of sequence modules for multi-domain proteins is well documented (Doolittle & Bork, 1993) and supports the idea that domains fold independently, too (Wu *et al*., 1994).

General geometrical specification of interacting protein surface areas was considered only in a few cases, e.g. for interfaces of subunits (Connolly, 1986) and for binding sites of small ligands (Peters *et al*., 1996), though the development of shape descriptors for database screening is under way (Good *et al*., 1995). Binding sites of small molecular substances, up to 2 kDa, are generally found to be small curved pockets (Peters *et al*., 1996), in contrast with interacting surfaces between two proteins, which are quite flat (Jones & Thornton, 1996). Until now the geometric properties of interacting pieces of secondary structures have not been described in detail.

In the analysis presented here we begin with an extensive study of molecular interfaces between secondary structures and between secondary structures and solvent. We started with a compilation of all such interfaces for a given set of protein structures. On the basis of their secondary structure we disassembled the proteins into molecular surface patches, which were further classified by their direct neighbor. These pairs of patches were deposited as interfaces in a data bank. Furthermore, a query system was developed to find similar molecular surfaces or interfaces, respectively.

For effective filtering, we derived properties of the interfaces describing their geometric and molecular (atomic) features.

This work will provide a method to examine and predict protein folding as a concerted docking process of pre-formed secondary structures. Moreover, it presents a further opportunity for rational design and recognition of binding sites of proteins using patches of known molecular surface areas as parts of a jigsaw puzzle.

Results and Discussion

The aim of our approach is the derivation of complementary molecular surface patches (MSP) useful for docking prediction within and between proteins. For these purposes it was necessary to develop a representative and adaptable data collection of experimentally observed complementary MSPs, including a unique procedure to dissect the total surface. This allows for combination of joined interfaces to build the complete docking or binding site. We analyzed relevant properties of the MSPs, and generated a retrieval system that permits the search for 3D similar interfaces in the database, including a fast algorithm to compare two MSPs at the atomic level. To demonstrate the efficiency of the data bank and its query system, some cases of similarity of MSPs are presented.

The elements of the data bank DIP/glossary

The motivation for creating the data bank was to obtain representative pairs of complementary surfaces (interfaces) from protein structures. Direct contact in this context means that atomic van der Waals (vdW) surfaces are closer than a given cut-off distance.

The basic elements of our dictionary of interfaces of proteins are molecular surface patches of protein substructures. Due to the fact that the exterior and the interior surfaces do not show any genuine ''starting point'', and that a particular substructure has more than one neighbor, a meaningful dissection of this continuous molecular surface area had to be introduced. This procedure corresponds to a widespread approach to disintegrate multiple problems into a number of pair problems. Therefore, we selected an unambiguous, reversible procedure to divide the protein structure, as well as the interface between protein and solvent, into patches. After the dissection of the molecular surface, the patches were stored in a data bank and were characterized. In the following section these steps are described in detail.

Data structure of the DIP

The basic contents of the DIP are sets of atoms defined by their membership of a distinct interface. For practical reasons, some information was adopted from the PDB. Further descriptors of surface area and atoms were held in separate files.

The database DIP consists of three types of cross-referenced files (see also Figure 1).

I. All MSPs are listed in a master file containing unambiguous ID numbers for: (1) the protein; (2) the interacting secondary structures (the patches in contact with solvent are handled in a comparable fashion); (3) the two MSPs building an interface.

II. Interface file for every PDB entry, which gives the following information (see also Appendix): compound; source; author; resolution according to the PDB; number of atoms, amino acid residues, SSEs including type and length; arrangement of hetero-atoms; atoms of every SSE are given in PDB notation, including the type of atom and the 3D coordinates taken directly from the PDB file. Their contacts to neighboring secondary structures are added to each atom to a maximum distance between vdW spheres of 2.8 Å. Thus a reconstruction of the protein as a jigsaw puzzle of interfaces is possible. The vdW volumes and solvent-excluded volumes are given for each atom (Goede *et al*., 1997).

III. The MSPs are classified according to their size. ID numbers are stored in special files (size files) sorted according their length, width and depth for intervals of 2 Å (from [2 Å, 2 Å, 2 Å] up to [56 Å, 42 Å, 26 Å]). In addition to the patch ID, the exact size (in all three dimensions) and the number of atoms of the patch are included as well as its atomic composition and shape.

Glossary

Secondary structural element (SSE)	Unit of secondary structure according to the algorithm of DSSP (Kabsch & Sander, 1983); only three types are used: α-helix, extended structure and coils
Non-protein elements	Ligands, substrates (hetero-atoms), explicit inner solvent, virtual outer solvent (described as continuum and considered as one large structure)
Cut-off	Allowed maximum distance between vdW spheres of an atomic contact ($\leqslant$2.8 Å)
Neighboring atoms (atoms in contact)	Two atoms (from different SSEs) within a distance less than the sum of their vdW radii and the cut-off value
Atomic surface area	Surface area defined according to Connolly (1983); surface around the protein that is not accessible to a solvent molecule
Inner Connolly surface	Surface enclosing cavities within proteins (larger than one water molecule)
Molecular surface patch (MSP)	Set of atoms of a given secondary structure that are in contact with atoms of another structure (secondary structural element, solvent, ligands)
External MSP	All atoms of one secondary structure that are in contact with the Connolly surface of the protein
Interface (internal)	Pair of MSPs from different structural elements in direct contact

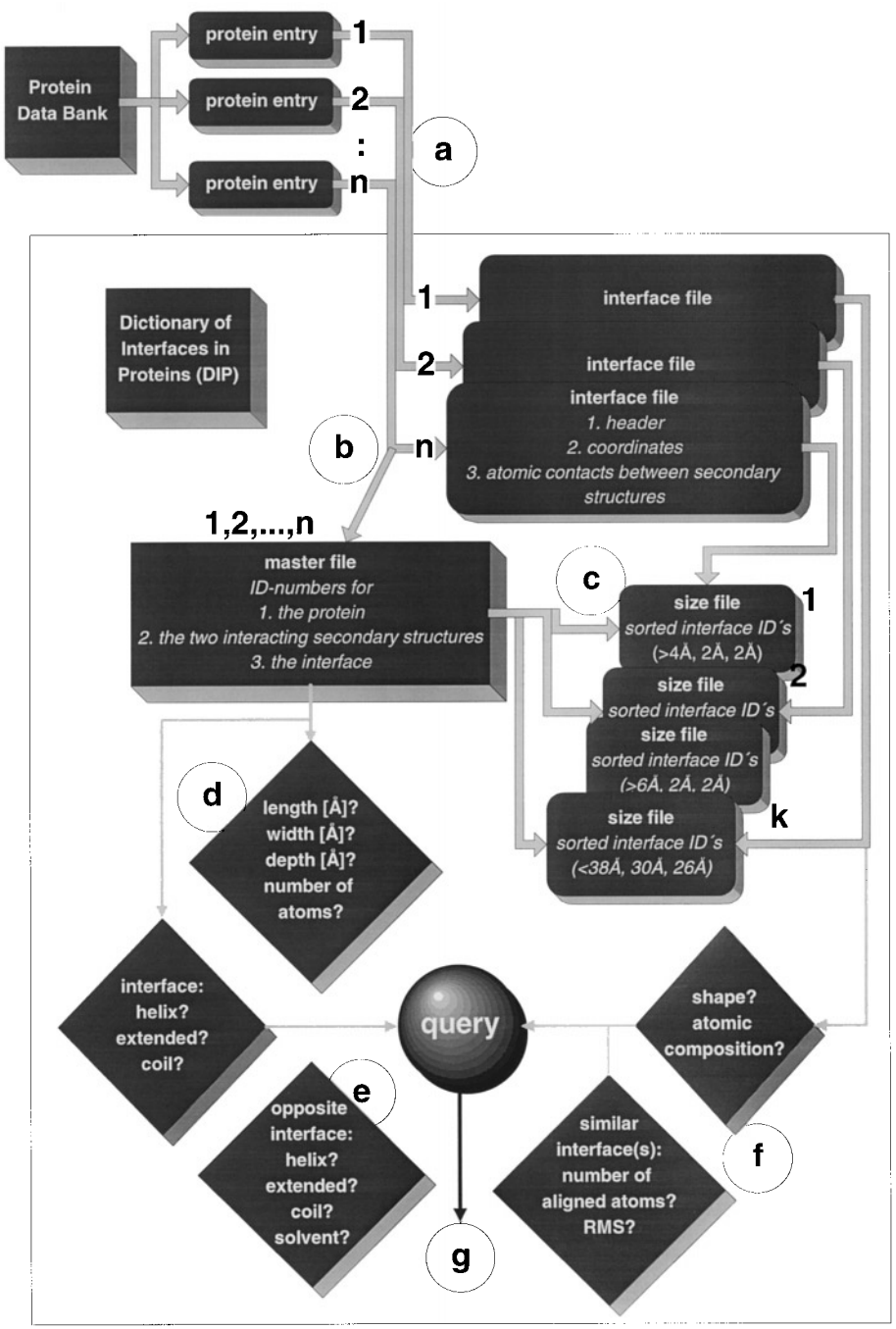

Figure 1. A flowchart showing data fluxes of the DIP. (a) Each protein from the PDB can be added to the DIP or special subsets (e.g. families of homologous proteins) can be selected. The interfaces between SSEs are stored in interface files. (b) For retrieval, a master file is generated containing unambiguous ID numbers for the access of each interface. (c) The access to subsets of interfaces is accelerated by presorted interface IDs according to distinct intervals of length, width and depth. Further information like number of atoms, atomic composition or shape are also stored in the size files. Query: The retrieval system can answer questions of the following type: find interfaces of a given size (e.g. longer than 10 Å, thinner than 4 Å, ...); find interfaces of a given type (e.g. helix to solvent); find interfaces similar to a given original (extracted from PDB or interface files). A combined search considering these constraints can be carried out as follows. (d) For restriction of the structural alignment procedure on patches with a distinct interval (±2 Å) of length (17 Å), width (11 Å), depth (7 Å) and number of atoms (55 to 65), size files are used to reduce the search space on 500 interfaces. (e) The type of searched interface will be found by the retrieval system scanning the size files, in this case helix. This further reduces the number of considered examples to 113. (f) The additional input for the fast sequence-independent structure alignment screening for similar patches is the required minimal number of aligned atoms (>50) at the permitted maximal rms value (<0.5 Å).

Query-results: (g) Looking for a helix-helix interface similar to a helix-patch (PDB code 1gca; 212-224) we found an example with 51 atoms and an rms value of 0.40 Å, which is illustrated in Figure 8.

Table 1. Characteristics of the database

Characteristic	Number
Number of proteins	351
Number of residues of the smallest protein considered	45
Number of residues of the largest monomer of a protein	680
Mean number of residues per protein	248
Total number of residues in helices/number of helices	24,694/2,176
Total number of residues in sheets/number of extended strands	20,207/3,822
Total number of residues in coiled regions/number of coiled regions	42,196/6,477
Total number of residues total/number of structural elements	87,097/12,475
Total number of atoms	670,000
Number of hetero-atom complexes bound to proteins	756

The given data structure is advantageous for several demands upon the DIP: speeding up the search for similar patches; accelerating the localization of MSPs with distinct properties (e.g. size, atomic composition); and reconstruction of complex binding sites that consist of more than one pair of MSPs. For this reason it is necessary to know which SSEs are neighbors (master file), and to maintain access to the original coordinates (interface file).

Characteristics of the current content of the DIP

A survey of characteristics for the 351 proteins included in the present data set is given in Table 1. With a cut-off value of 2.5 Å (see below), the molecular surface area of any particular SSE (total 12,475) on average contains 13 (partially overlapping) patches. Partially overlapping means that one distinct atom can be the neighbor for atoms from different structural elements, and thus belong to different patches. The 160,000 interfaces in our database are located at the surface of 2176 helices, 3822 sheets and 6477 coils. The average length of sequentially consecutive structural elements in the database is about seven amino acid residues, which corresponds to the average size of coils (between 1 and 86 residues). The average length of helices is clearly larger (about 11 residues, ranging from 4 to 42), that of single strands of sheets is somewhat smaller (mean 5.0, ranging from 2 to 19). Therefore, a protein of 250 amino acid residues can be dissected, on average, into 35 SSEs. Even small proteins show relatively large patches in repetitive SSEs. A protein with more than 100 residues contains at least two, usually more than five, large SSEs; i.e. helix longer than 11 residues, extended strand larger than four residues, coiled segments longer than seven residues. Each of these larger elements is involved, on average, in eight interfaces consisting of more than four atoms on each side (at cut-off 2.5 Å), five of which consist of more than ten atoms and at least one with more than 30 atoms on each side. Hence, half of the atomic positions (e.g. 500) of a protein (containing about 1000 heavy-atoms) can be reconstituted directly by positioning merely the five largest pairs of patches from the five largest SSEs (each with more than 20 atoms). The symmetric matrix containing the number of contacts between SSEs (Table 2) reflects predominantly size and number of the different types of secondary structural elements.

The influence of the allowed tolerance in the contact distance

In proteins, the packing of atoms is almost perfect (Richards, 1974, 1977). In this respect, the distance between vdW spheres for non-covalently linked neighboring atoms should generally be close to zero. But in folded proteins defects often occur. The larger the protein, the larger the number and size of the defects (Williams *et al.*, 1994; Hubbard *et al.*, 1994; Hubbard & Argos, 1995). Furthermore, the atomic coordinates are assumed to show an experimental error well below 0.5 Å (Thornton *et al.*, 1990). Unfortunately, the distinct values of vdW radii, at least for united atoms, are not unambiguously defined. As a consequence, the distance between neighboring atoms can be larger (and even smaller) than the sum of the vdW radii. Therefore, to decide whether two atoms are in con-

Table 2. Number of molecular surface patches in 351 proteins

	α-Helix	Extended structure	Coiled structure	Hetero-compound	Solvent
α-Helix	11,102	11,096	22,490	2160	2139
Extended structure		33,446	44,606	1918	3745
Coiled structure			36,606	3640	6350
Hetero-compound					534

Each observation is counted if at least one atom of a distinct SSE has a distance smaller than the cut-off value of 2.5 Å to any atom of the other element. A list of proteins is given in Materials and Methods. The secondary structure is defined according to DSSP: α-helix, extended structure (one strand of a parallel or anti-parallel β-sheet), coiled structure (not helix, not β-sheet), heterocompounds bound to the protein (non-protein atoms), solvent (exterior, described as continuum).

tact, deviations from their ideal contact distance have to be tolerated. This tolerance value (allowed distance between vdW surfaces of atoms) is called cut-off distance and is utilized in our definition of interfaces (see also Determination of neighborhood in Materials and Methods).

Obviously, the total number of interfaces, as well as the total number of atoms in interfaces, increases with the chosen cut-off value. This is valid for both internal contacts and contacts to solvent. As a result, the size of interfaces and their atomic composition are also cut-off-dependent (see Figure 2). The number of atoms in interfaces increases greatly up to a cut-off value of 1.5 Å. Upon further increase of the cut-off value, the size of the patches increases more slowly. In most cases, the shape of contact areas is changed only slightly with rising cut-off (see Size and shape of molecular surface patches … , below).

The number of interfaces per single SSE also increases with increasing cut-off (see Figure 2), but the size of newly assigned neighboring structures diminishes dramatically at higher cut-off values. At higher cut-off values (e.g. 2.5 Å) only small interfaces (on average consisting of three atoms) were added to the list of neighbors of a given secondary structure. Patches appearing for the first time at a cut-off value of 1.5 Å are, on average, five atoms large and those appearing at 0.5 Å consist of eight atoms. Only 10% of later-appearing interfaces contain more than ten atoms at a cut-off of 2.5 Å.

The influence of the cut-off distance on amino acid prevalence is less striking than its influence on atomic composition (data not shown).

The appropriate choice of cut-off distance

The necessary consideration of the cut-off distance in the definition of interfaces makes it an additional parameter showing influence on different properties of the MSPs. Keeping in mind that distinct types of interfaces at lower cut-off values show significant differences in atomic composition, and that geometric features will be more pronounced at higher cut-off values due to the larger number of atoms, cut-off values between 1.0 Å and 2.0 Å appear advantageous. Cut-off values larger than the diameter of a water molecule are not meaningful, because SSEs would be evolved from second neighbors. Because no general optimal choice of cut-off value exists, the minimum cut-off value for any atomic contact was stored.

Contact between secondary structural elements

There is a strong linear relationship (correlation coefficients of $r \approx 0.9$) between the number of amino acid residues per unit and the number of adjacent interfaces, which is only parallel-shifted for different values of cut-off (lowest line in Figure 3). Considering the mean number of neighboring residues instead of neighboring elements, the linearity is even stronger. The slope of all four lines is similar: one additional neighbor per three residues in helices and coils, but extended elements are found to have significantly more neighbors per residue (30 to 50%) than helix or coil (two residues in extended structure, one additional neighbor).

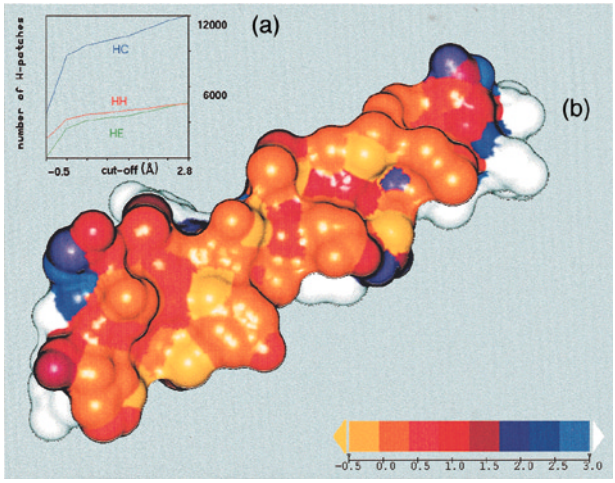

Figure 2. Visualization of the cut-off dependence of the expansion of a helical patch. (a) Helix A (residues 14 to 41) of ferritin (PDB code 1FHA) as seen from its neighboring helix B (residues 50 to 76). The color code shows the cut-off at which atomic contacts occur. (b) (Inset upper left) The rise of the number of helix-patches with cut-off for the database of 351 proteins.

Contact between secondary structural elements and solvent

Generally, the interface area to solvent includes the part of the protein that interacts with ligands like substrate, coenzymes, other proteins, etc. For docking analysis, an adequate description of this external surface area analogous to internal surface area is required. But for solvent-accessible surface area the molecular neighbors are often not given explicitly. Only a few coordinates of the solvent or ligands are listed in the PDB. To introduce the cut-off dependence of membership of solvent-accessible atoms to external patches, we had to develop a comparable definition for this case. For practical reasons, we prefer the molecular surface area according to Connolly (1983) instead of artificially soaked proteins (Eisenhaber & Argos, 1996). The computation of the Connolly surface does not require coordinates of surrounding solvent molecules but searches for the closest position of water molecules at each point and calculates the distance to protein atoms. A graphic explanation of the procedure to estimate contact to solvent is presented in Figure 4(b).

At a cut-off value of 0.0 Å we found only 350 SSEs (2.8%) that were completely covered. Almost two-thirds of these completely buried parts are sheets, a smaller amount are shorter coils and there are only three helices with more than ten amino

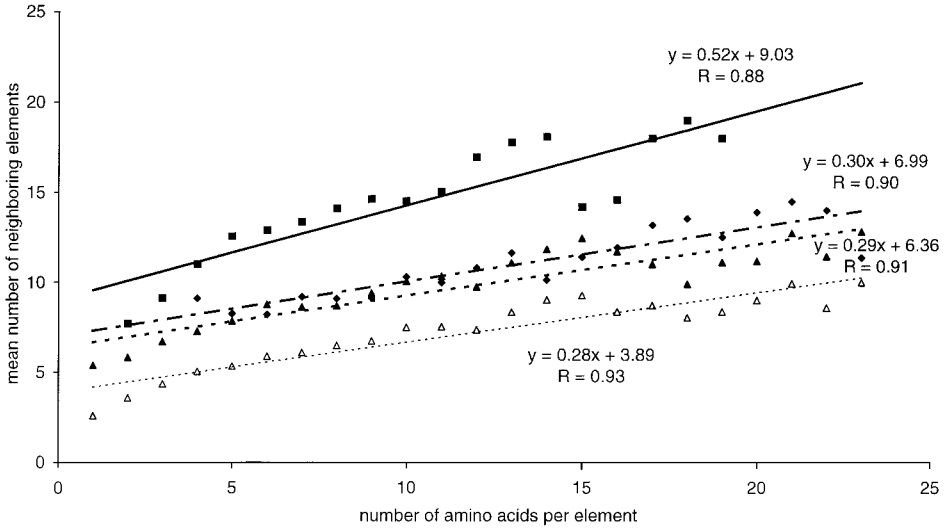

Figure 3. The number of neighboring SSEs in dependence on the length of the original element. The three upper lines are for contacts at a cut-off value of 2.5 Å. The bold line with filled squares (■) marks the sheets; the dash-dot line with the diamond symbol (◆) denotes helical contacts; the bold broken line with filled triangles (▲) stands for coil; the dotted line and the open triangles (△) designate the contacts of coils at a cut-off value of 0.0 Å. The equation and correlation coefficient (R) of the regression lines are given.

acid residues. At a cut-off value of 2.5 Å only 241 SSEs remain inaccessible for the solvent. The total number of external patches is smaller than the number of patches that are in contact with other secondary structures. In contrast, the mean size of external patches is larger. These two findings were not unexpected, because solvent-accessible SSEs have the solvent as one continuous neighboring structure. However, due to the definition of patches it is possible that one patch falls into two or more disconnected pieces (atomic distance larger than twice cut-off). The disruption is caused by other parts of the protein partially shielding the particular secondary structure. We checked the DIP for such cases and found only 1.7% of all patches (4% of exterior patches) interrupted.

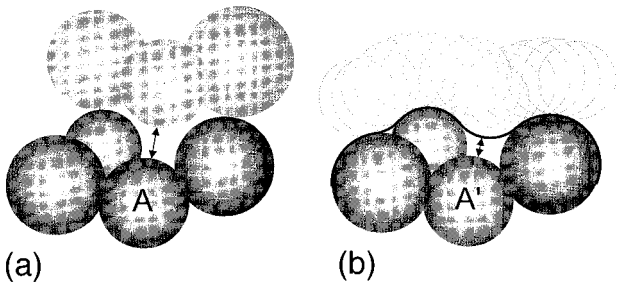

Figure 4. Comparison of cut-off definitions for interior and exterior atomic contacts. (a) The cut-off distances are marked by double arrows. The first approach is used for interior contacts and (b) the second (exterior contact) in the absence of an opposite SSE replaced by a virtual solvent. (a) Shortest distance between atom A' from one secondary structural element (four dark shaded atoms) and the neighboring unit (three lighter shaded spheres). (b) Shortest distance between atom A' from one secondary structural element (four dark shaded atoms) and the Connolly surface (bold line). The virtual solvent molecules are indicated by thin circles.

Size and shape of molecular surface patches as components of interfaces between secondary structural elements

The vdW surface of SSEs is convoluted and, according to our definition, composed of several MSPs. An understanding of general geometric properties of the participating patches would be useful both for the spatial description of docking sites and for an effective query system. Therefore we checked all types of interfaces for common geometric features. The centers of the atoms in a mean MSP are distributed in a rectangular parallelepiped with edge lengths of 9 Å, 5 Å and 3 Å. Comparing the three dimensions of all patches (Figure 5(a) and (b)), we found that one dimension is significantly smaller than the remaining two. Typically, 70% of the atoms of MSPs lie in a layer ±1 Å to their least-squares plane. Due to the larger extent of solvent-accessible MSPs and the shape of globular proteins, the deviation from planarity is slightly larger too. On average, only half of the atoms of the exterior patches have a distance of less than 1.0 Å to the corresponding least-squares plane.

Generally, the patches with more than ten atoms are approximately flat parallelepipeds with a length to width to depth ratio of about 3:2:1, independent of the cut-off used and the number of atoms related to the interface ($n_{atom} > 10$), respectively. This is valid for interior as well as for exterior patches.

Considering the local geometry of MSPs, the question arose of whether a particular patch is dominated by consecutively covalently linked atoms. As found earlier for domain interfaces

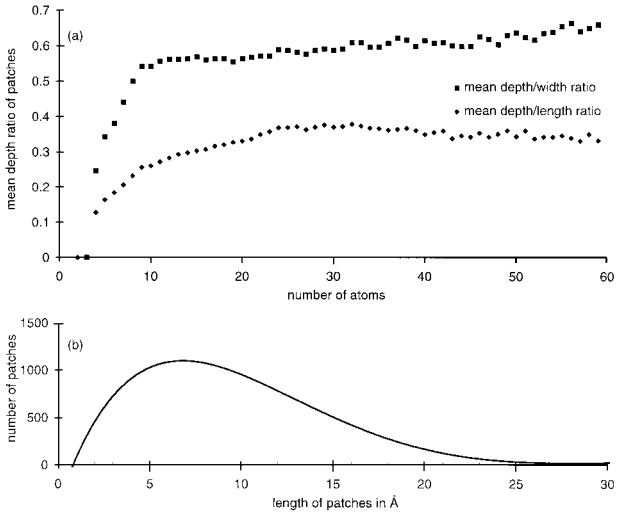

Figure 5. The three dimensions of MSPs. (a) Ratio between mean depth and length for MSPs with the same number of atoms. (b) Distribution of the length of 150,000 patches from the DIP.

(Argos, 1988), the majority of interfaces consist of only a few covalently linked subareas. A mean patch with 15 atoms is typically made up of three to five groups of covalently linked atoms. The mean number of covalently linked atoms in one patch is about five and is independent of size and type of SSE. In only a few cases is one complete side-chain the basis of the patch.

The largest helix-helix patch is observed in a lipoprotein (PDB code: 1LPE), with about 52 Å length, 14 Å width and 9 Å depth. The interface containing the highest number of atoms is a coil-coil contact in a DNA-binding protein (PDB code: 2GN5) with 172 atoms on one side and 144 atoms within the other patch. At a cut-off of 2.5 Å, the longest extended-solvent patch contains 150 atoms consisting of 18 amino acid residues. In accordance with the greater size of helices, the largest helix-solvent patch contains about 300 atoms. The structure with the most extensive contact to the solvent is a large coil in cytochrome (PDB code 1CY3; 637 atomic contacts).

The relative orientation of molecular surface patches in interfaces

Generally, in docking simulations six degrees of freedom have to be considered. To speed up structural alignments and docking computer experiments it is advantageous to define a general shape and relative orientation of molecular patches. As shown above, the MSPs can be described as relatively flat bodies. Due to this property it is possible to determine a normal line of this rectangular parallelepiped solid. Considering the distance of the centers of mass of particular pairs of patches we observe only a small variation of this distance around the expected value (the sum of mean vdW radii plus particular chosen cut-off distance; see Figure 6(a)). Furthermore, for different cut-off values the distribution of the angle between the normal lines of the interacting patches was estimated (Figure 6(b) and (c)). Considering patches with more than ten atoms on either side, at cut-off = 0.0 Å the angle between the axes ranges from zero to 30° in about 90% of the interfaces. The restricted relative shift of patches and the small range of relative torsion results in good starting points for docking simulations.

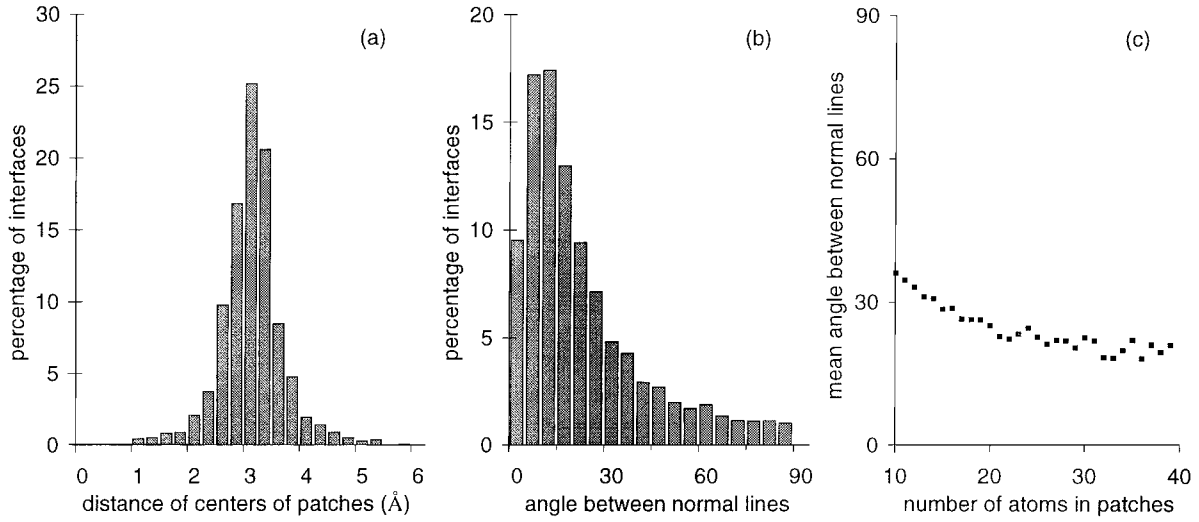

Figure 6. Relative position of the interacting patches. Only interfaces with at least six atoms on each side were considered. The size is given by the total number of atoms in both patches, the angles in degrees. The patches were evaluated at a cut-off value of 0.0 Å. (a) Frequency distribution of the distance between the centers of mass amongst opposite MSPs. (b) Distribution of the angle between the normal lines of both sides of interfaces (MSPs). (c) Mean angle between normal lines in dependence on the size of the patches.

Packing of molecular surface patches

Local atomic density is calculated as the ratio of vdW volume to the corresponding Voronoi volume (Richards, 1979). In proteins, the mean atomic density is quite high (0.74) and falls within a narrow range (0.70 to 0.78) in spite of the substantial variation of density observed within different parts of proteins (smaller than 0.60; larger than 0.90; see Figure 7(a)), as reported earlier by Richards (1974, 1977).

On the other hand, about 12% of atoms are closer to other atoms than would be allowed for vdW contacts, which we take into account by slightly negative cut-off values (see Figure 2). Most of these close contacts occur for hydrogen bonded atoms and, consequently, only a small amount of the non-covalent contacts in the database are actually too close.

As a first approximation, the local density between different elements can be considered as a measure of complementarity. The questions arose of whether the packing in different types of interfaces deviates from the mean and how the different types of patches contribute to the packing. To address these questions, we examined the local density in the MSPs of proteins with the Voronoi cell method (Goede *et al.*, 1997). For the mean local density, we observed no significant dependence on the number of neighboring patches but the distribution becomes sharper with increasing number of neighbors (data not shown). The local atomic density depends particularly on the type of secondary structure and its neighbor. We find a significantly higher packing density in ladders of β-sheets (mean 0.72) than in coils (mean 0.63), which yields a sharper peak in Figure 7(b) than in Figure 7(c). Surprisingly, the corresponding density in the neighboring helices differs with the same trend: in contacts with ladders of β-sheets the density is higher (0.70) than in contacts with coiled structure (0.66; see Figure 7(b) and (c)).

Atomic composition and amino acid preferences of molecular surface patches

If the geometry and the physiochemical composition show distinctive properties, the concept of DIP will be useful for recognition of distinct types of MSPs on the surfaces of given 3D structures of SSEs.

The composition of the different types of interfaces in the database reflects known principles of protein folding: the interior surface areas of SSEs show more apolar atoms and apolar amino acid residues than solvent-oriented parts. In a ranking of all types of patches according to the content of oxygen and nitrogen atoms, the three patches oriented toward the solvent take the first three positions.

About two million atomic contacts were analyzed to find typical arrangements (e.g. for an EH patch in comparison with an EC patch). The first remarkable point is the main-chain to side-chain proportion, which differs dramatically for these patches at cut-off value 0.0 Å: 12% main-chain atoms in EH contacts; 25% main-chain atoms in EC contacts. Such effects are partly reflected in the polarity of the particular patches: totally EC-MSPs exhibit double the number of polar atoms compared

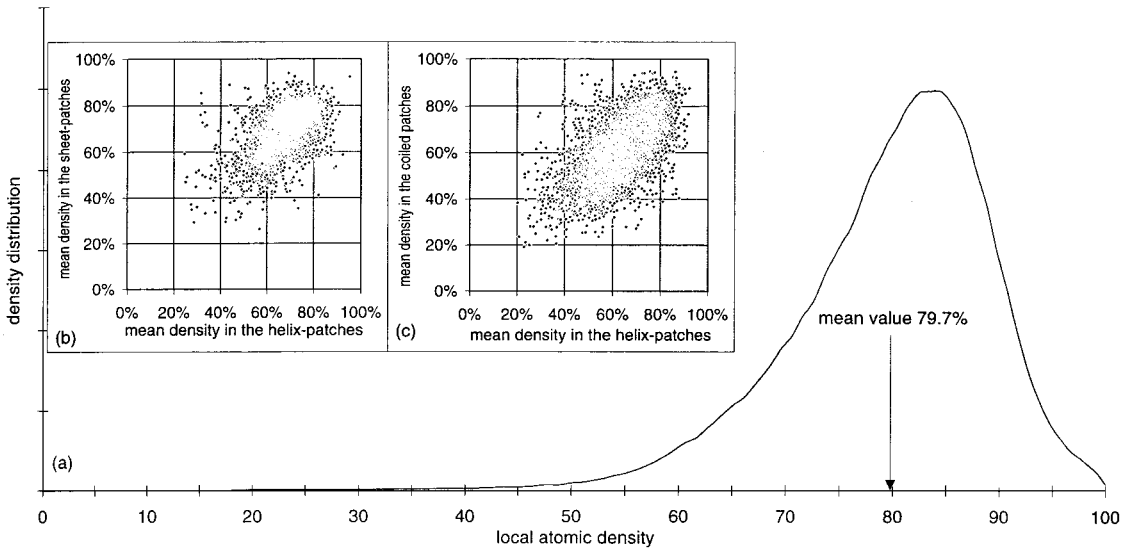

Figure 7. Comparison of local density distributions in MSPs. (a) Density distribution at the surfaces of buried SSEs. (b) Local atomic density of MSPs localized on sheets in dependence on the corresponding density of MSPs localized on helices for sheet-helix contacts. (c) Local atomic density of MSPs localized on coils in dependence on the corresponding helix-density for coil-helix contacts.

Table 3. Portion of polar atoms in molecular surface patches in the 351 proteins of the DIP

	Helix	Extended structure	Coiled structure	Hetero-compound	Solvent
Helix	16	13	31	24	34
Extended structure	14	33[a]	27	30	36
Coiled structure	31	28	31	31	38

Polarity is defined as percentage of oxygen and nitrogen atoms in the particular interface.
[a] 82% main-chain atoms.

to EH-MSPs (see Table 3). Interestingly, the average HE patch shows exactly the same composition as its opposite average EH patch, while EC and CE patches (30% main-chain atoms) clearly differ.

At lower cut-off values, larger differences concerning atomic composition occur between the different types of MSPs. Mainly, differences in the atomic composition due to variance in amino acid composition remain visible at higher cut-off values. Considering distinct types of SSEs in detail, we observe different cut-off dependencies of atomic composition for particular contact types. Such detailed considerations including cut-off dependence were carried out for all types of patches and will be published elsewhere. In consequence, regions of contact to a particular type of SSE should become predictable for a given SSE.

The database DIP includes a total of 87,000 amino acid residues involved in interfaces. We calculated propensities for each type of amino acid and secondary structure (data not shown) and compared them with the propensities presented by Swindells *et al*. (1995). The propensity of an amino acid in a particular secondary structure is the ratio between the proportion of the amino acid in this type of secondary structure and that in the complete databank. Values larger than 1 can be found, especially for Ala, Glu, Leu, Met and Gln in helices, and Phe, Ile, Val and Tyr in sheets, showing that the appearance of these amino acids in the corresponding secondary structures is above average. Values less than 1 for both helix and sheet (Asp, Gly, Asn, Pro and Ser) are indications for their well-known frequent occurrence in coils. Propensities calculated here and by Swindells *et al*. (1995) show good agreement (correlation coefficient 0.975), indicating a representative selection of proteins in our database.

Using the DIP, we are able to estimate the propensities for amino acids occurring in definite pieces of SSEs, e.g. helix (H) in contact with other SSEs (e.g. HE, HC and HH). The propensity of an amino acid in a particular type of MSP is the ratio between the proportion of this amino acid in these MSPs and its proportion in the corresponding type of secondary structure (see the legend to Table 4).

In this way, propensities for particular regions of helices express deviations from general helix propensities and values around 1.0 indicate agreement with those for total helices. We find one-sixth of the propensities deviating more than 25% and a few more than 50% (see Table 4). These data complete the picture that arose from the atomic composition. If we compare propensities, e.g. of Gln for EH (0.56) and EC patches (1.01), strong differences occur, which will be valuable for structure predictions. Charged residues are clearly solvent-oriented.

Summing the results of atomic composition and amino acid prevalence of MSPs, the prediction of the probable neighbors of a patch should become a solvable problem.

Geometrical similarities between patches at the atomic level

Typically, the following task arises in a number of modeling studies: to create a complementary binding partner (substrate, inhibitor, etc.) for a given active site (enzyme, receptor, etc.). Using the DIP, we can rapidly search for MSPs similar to the binding pocket with the additional condition of an existing opposite. These opposites can be used as leading structures for ligands. In this way we transform the problem of finding complementarity into a retrieval for similarity. The existing retrieval system of the DIP, including an automatic superposition procedure, allows the search for similar MSPs. Here, we present preliminary results of similar patches in distantly related proteins. For evaluation of the alignment procedure, distinctly related proteins were included in the data set: e.g. arabinose-binding protein (PDB code 1ABE) and galactose-binding protein (PDB code 1GCA). The sequence identity of these proteins is below 25% but, as demonstrated in Figure 8, the interhelical MSPs are well conserved. The detection of such sequence-independent similarities in arbitrary proteins can be carried out in a straightforward manner with the retrieval system of DIP as described in Figure 1.

A surprising result of a large-scale search was the detection of structurally analogous interfaces with up to 50 atoms in different types of SSEs. We were able to superimpose up to 24 of 30 atoms from such MSPs with an rms value smaller than 0.5 Å. On average, more than 50% of these atoms coincide in basic atomic properties like partial charge and hydrophobicity. Even examples with reversed chain direction occur (Preißner *et al*., 1997).

Table 4. Amino acid propensities $\langle P \rangle$ for MSPs in 351 proteins

$\langle P \rangle$AA	A Ala	C Cys	D Asp	E Glu	F Phe	G Gly	H His	I Ile	K Lys	L Leu	M Met	N Asn	P Pro	Q Gln	R Arg	S Ser	T Thr	V Val	W Trp	Y Tyr
$\langle P \rangle_{HH}$			0.79	0.85	1.21			1.17	0.87	1.22	1.18	0.84		0.89		0.84		1.13		1.15
$\langle P \rangle_{HE}$		**1.28**	**0.75**	**0.75**	**1.30**		0.76	**1.43**	**0.67**	1.16	**1.32**	0.86	**0.66**	0.88	0.88			1.19	**1.49**	**1.36**
$\langle P \rangle_{Hsolv}$						0.88														
$\langle P \rangle_{CH}$				0.89	**1.27**		1.16	1.19		**1.29**	**1.35**					0.88			1.18	1.14
$\langle P \rangle_{CC}$		1.11			1.14	0.89					1.12							1.11	1.18	1.19
$\langle P \rangle_{CE}$		1.12						1.14	0.85									1.16	1.15	
$\langle P \rangle_{EH}$		0.86	**0.75**	**0.73**	**1.30**	0.85		**1.34**	**0.72**	1.20	**1.32**	0.83		**0.56**	0.88	**0.68**	0.77	1.17	1.14	1.15
$\langle P \rangle_{Esolv}$	0.89			1.18		0.83	1.15		1.21	0.88		1.14		1.14	1.19					1.11

Amino acids are given in both the one-letter and the three-letter code in the first row. The abbreviations in the indices of the first column follow the conventions of the secondary structure notation in this paper. $\langle P \rangle_{Hsolv}$ means amino acid preference for solvent-directed helix-patches. Details concerning estimation are described in Materials and Methods. For clarity, only values differing more than 10% from expectation are given. Values deviating more than 25% are in boldface and small values are additionally underlined.

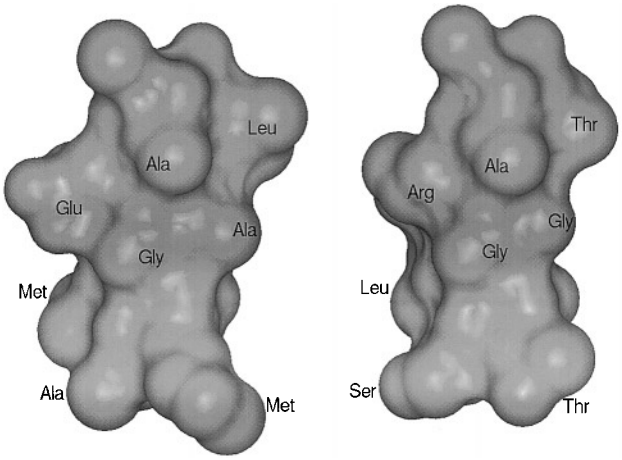

Figure 8. Example of similarity between MSPs from related proteins. Superposition of the inter-helical MSP built by residues 212 to 224 of the galactose-binding protein (PDB code 1GCA; left) and one side (residues 206 to 218) of the interface between two helices in arabinose-binding protein (PDB code 1ABE; right). The overall as well as the local sequence identity of these proteins is below 25%. A possible search strategy is sketched in Figure 1. In all, 51 of 60 atoms could be superimposed with an rms deviation of about 0.4 Å. Only 4% of the aligned atoms were of different type.

An exhaustive search for similar patches in homologous proteins will be presented elsewhere.

General discussion and future directions

The DIP and its retrieval system were designed as a tool for protein folding and docking simulations, respectively. Defining molecular patches as complementary surfaces of structural elements, we get a reasonable first partitioning of the continuous, endless surface area of complex protein molecules and their ligands. The DIP represents a large and comprehensive data set of inter- and intramolecular patches. The database includes a reorganized, preliminarily classified form of information derived from the Brookhaven protein structure Data Bank (PDB), which will be helpful for more detailed analysis of molecular architecture of interacting molecular surfaces. This data bank is supplementary to other hierarchical classifications of protein structure on the basis of secondary structural elements (Orengo *et al.*, 1997). The classification by simple geometrical and physicochemical features of MSPs can be useful for prediction and evaluation of any interacting molecular pair, e.g. between secondary structures or between protein and ligand. Future considerations will be undertaken in two directions: structural prediction of proteins starting from primary structure; and a search for ligands of proteins non-covalently bound at their molecular surface, respectively.

First results concerning atomic composition and amino acid preferences in the different types of interfaces between secondary structures and their neighbors give hope that the inverse process, predicting the neighbor of a given patch (using its properties), will be possible with sufficient probability. The solution of the docking problem is mathematically very complex (Kuhl *et al.*, 1984). Consequently, it would be reasonable to learn from available data of known structures of complementary molecular surface patches of proteins.

Because the DIP contains information about all the neighbors of a particular patch, the reconstruction of complex active (docking) sites is possible by means of combination of molecular patches. The reconstruction can be done at the level of simple geometric bodies resulting from intersecting planes (low resolution) as well as at an atomic level (high resolution).

Summarizing, the dictionary of interfaces in proteins will be a basis for better understanding of the geometrical and physiochemical complementarity as well as a tool for prediction of docking processes.

Materials and Methods

Database

The starting point for the DIP is a set of protein structures each given as a compilation of its atomic coordinates. The content of the database DIP can easily be adapted to the problem concerned (e.g. exclusively members of a family of homologous proteins, or proteins binding certain ligands). The generation of the database in each case is guided by a list of proteins to be considered. For general purposes, a representative data set with low redundancy (<10%) was chosen (listed below using the 4-character accession code from the PDB).

135L; 155C; 1AAK; 1AAN; 1AAW; 1ABA; 1ABE; 1ABK; 1ABM; 1ACB; 1ACE; 1ACX; 1ADS; 1AK3; 1AKE; 1ALA; 1ALB; 1ALC; 1ALD; 1APM; 1ARB; 1ARC; 1ATN; 1AZU; 1BBC; 1BBH; 1BGC; 1BGE; 1BIA; 1BOV; 1BP2; 1BSR; 1BTC; 1BTI; 1C2R; 1CAD; 1CBN; 1CC5; 1CCR; 1CD4; 1CDT; 1CGI; 1CGJ; 1CHO; 1CMB; 1CMS; 1COL; 1CP4; 1CSC; 1CSE; 1CTH; 1CY3; 1CYC; 1DFN; 1DMB; 1DR1; 1DRF; 1DRI; 1EAF; 1ECA; 1END; 1EST; 1EZM; 1F3G; 1FAS; 1FDH; 1FDL; 1FDX; 1FHA; 1FIA; 1FKF; 1FLV; 1FNR; 1FUS; 1FVC; 1FX1; 1FXA; 1FXD; 1FXI; 1GAL; 1GCA; 1GCT; 1GHL; 1GKY; 1GLT; 1GLY; 1GMF; 1GOX; 1GP1; 1GPR; 1HDS; 1HEL; 1HHL; 1HLE; 1HNE; 1HOE; 1HPT; 1HRH; 1HSB; 1IFB; 1IGM; 1IPD; 1ITH; 1LDM; 1LEC; 1LH2; 1LLA; 1LLD; 1LPE; 1LTE; 1LVL; 1MBA; 1MBC; 1MBD; 1MBS; 1MEE; 1MNS; 1MRR; 1MSB; 1MUP; 1MVP; 1MYG; 1NCO; 1NDK; 1NN2; 1NPC; 1NPX; 1OFV; 1OMD; 1OMF; 1OVB; 1P12; 1PAL; 1PAZ; 1PEK; 1PGD; 1PGX; 1PI2; 1PII; 1PK4; 1PLC; 1POA; 1POC; 1POD; 1PP2; 1PPA; 1PPF; 1PPG; 1PPL; 1PPM; 1PPO; 1PTS; 1R69; 1RAT; 1RBP; 1RCB; 1RDG; 1RDS; 1REI; 1RHD; 1RNB; 1RNE; 1ROP; 1RPE; 1RTC; 1RTP; 1RVE; 1SBP; 1SDY; 1SGT; 1SHA; 1SHF; 1SIM; 1SMR; 1TEC; 1TEN; 1TFD; 1TFG; 1TGS; 1THB; 1THG; 1TIE; 1TON; 1TPK; 1TRB; 1TRM; 1TTA; 1UBQ; 1UTG; 1VAA; 1WSY; 1YAT; 1YCC; 1YEA; 21BI; 256B; 2AAA; 2ACH; 2ACT; 2ALP; 2APR; 2AZA; 2AZU; 2BP2; 2CAB; 2CCY; 2CDV; 2CI2; 2CMD; 2CNA; 2CPL; 2CPP; 2CRO; 2CTS; 2CYP; 2ER7; 2FB4; 2FCR; 2FKE; 2FXB; 2GBP; 2GCR; 2GN5; 2HAD; 2HBG; 2HMB; 2HMQ; 2HPR;

2IMM; 2LAL; 2LBP; 2LIV; 2LTN; 2MCG; 2MCM; 2MHB; 2NN9; 2OR1; 2PF1; 2PF2; 2PIA; 2PKA; 2PLT; 2POR; 2PRK; 2PTC; 2RAT; 2REB; 2RHE; 2RN2; 2RSP; 2SGA; 2SN3; 2SNI; 2SNS; 2SOD; 2TGA; 2TGP; 2TRX; 2TS1; 2TSC; 2WRP; 351C; 3ADK; 3APP; 3APR; 3B5C; 3BLM; 3C2C; 3CHY; 3CLA; 3CLN; 3CPA; 3CRO; 3CYT; 3DFR; 3DNI; 3EBX; 3FXC; 3GAP; 3GBP; 3GRS; 3IL8; 3LZM; 3PRK; 3PSG; 3RAT; 3RP2; 3RUB; 3SC2; 3SDP; 3SGB; 3TGL; 3TLN; 3XIS; 4BLM; 4BP2; 4CLN; 4CPA; 4CPV; 4DFR; 4ENL; 4FGF; 4FXN; 4GST; 4HTC; 4ICB; 4P2P; 4PEP; 4PFK; 4PTP; 4RAT; 4RXN; 4SBV; 4SDH; 4TMS; 4TNC; 5ADH; 5APR; 5CHA; 5CPA; 5CYT; 5FD1; 5HVP; 5ICD; 5P21; 5PAL; 5PTI; 5RAT; 5TIM; 5TNC; 6FAB; 6LDH; 6RAT; 6RLX; 6RXN; 6TAA; 6XIA; 7FAB; 7LPR; 7PCY; 7RAT; 7RSA; 8DFR; 8GCH; 8RAT; 8RUB; 8RXN; 9PAP; 9RAT; 9RNT.

The 351 sets of coordinates were obtained from the Brookhaven Protein Data Bank (Bernstein *et al*., 1977). The selection was done according to the following criteria. (i) Well-resolved structures are included. (Resolution better than 2.5 Å; lower resolution was accepted only in special cases like virus capsids, where errors in atomic positions are reduced by symmetry of the molecular complex.) (ii) If more than one entry exists for a distinct protein, the best-resolved was selected. For comparative reasons, for a few cases with well-resolved structures with and without substrate (inhibitor, ligand) both (three) structures were included in the data set. For cases with identical chains in the asymmetric cell without a large contact region, only one chain was used in our calculations.

To check that the database selected is representative, we compared it with several other sets and their properties. Comparing our data set with the ''25%-identity list'' by Hobohm & Sander (1994), we can point out that the DIP covers a comparable number of proteins (365) of about the same size (on average 213 amino acid residues). The content of coiled residues is equal in the databases, while in our set of proteins the helix content is 10 % lower in favor of the sheet content.

Methods

The first step in preparing the data bank DIP involved dissecting each protein into structural elements (see Definition of structural elements of proteins, above). The contacts between those elements were then determined (see Determination of neighborhood, above). For this purpose, accessible and Connolly surfaces were calculated (see Atomic surface areas, above). The relevant information of the PDB entries, as well as derived properties of the MSPs, like size, composition, contacts or density, were stored.

Definition of structural elements of proteins

The protein structure can be divided into a pure protein (or peptide) constituent built exclusively of amino acid residues and, on the other hand, a non-protein part defined in the PDB as a hetero-compound. In our analysis, the non-protein constituent includes organic compounds like prosthetic groups, substrates, inhibitors and activators, inner water molecules and ions. Distinct water molecules located at the outer protein surface are ignored. The external solvent is handled as a continuum.

The protein structure itself was dissected into structural elements on the basis of secondary structure. There is a variety of methods to assign the location of structural elements of proteins (Colloc'h *et al*., 1993). The assessment of secondary structure is an unambiguous procedure, but known algorithms are not free of artifacts (Colloc'h *et al*., 1993). To be comparable with other studies we chose the common method of the DSSP program written by Kabsch & Sander (1983) to assign helical segments, H (4-helix = α-helix), and extended β-strands participating in β-ladder, E. All other segments are summarized under coil, C. The most significant artifact of the Kabsch-Sander algorithm is the high frequency of short helices assigned in other methods as coil (Colloc'h *et al*., 1993). Because the secondary structure assignment is used only to dissect the protein structure into appropriate pieces, and the knowledge for patching up of neighboring units exists in the data bank, a possible misassignment does not influence our results substantially.

The non-protein structural elements consist of a set of atoms in the PDB listed as hetero-atoms.

Prosthetic groups and other bound ligands are handled like the SSEs. All hetero-atoms covalently linked to each other were grouped into one identity.

Determination of neighborhood

Within a heap of tightly packed spheres neither the determination of neighborhood nor the definition of borders between MSPs can be considered as unambiguous. A stringent contact criterion (vdW spheres touching each other) results in small and disconnected interfaces due to small packing defects. On the other hand, allowed distances between neighbors larger than the diameter of an atom lead to apparent neighbors, separated most probably by an intervening water molecule or by another part of the protein molecule. Therefore, the distances between the vdW spheres smaller than 2.8 Å (diameter of the smallest heavy-atom) are to be held in the database. Contacts between any atom and ''other elements'' are taken into account. Such other elements could be: SSEs of the same or another peptide chain that do not contain the given atom; external solvent described as the outer Connolly surface; internal solvent, meaning cavities in the protein larger than one water molecule (inner Connolly surface); hetero-atomic compounds (non-protein structural elements).

Atom radii, slightly influencing the contact definition, are used according to Stouten *et al*. (1993). There are significant differences (up to 0.28 Å) between vdW radii used in various analyses and simulations (e.g. see Chothia, 1975; Brooks *et al*., 1983; Brünger *et al*., 1987; Stouten *et al*., 1993). While using the DIP for different purposes, we realized that the cut-off range of particular interest is between 0.5 Å and 2.0 Å, and that it is appropriate to store all interatomic distances up to 2.8 Å, covering the influence of different sets of atomic radii.

Determination of molecular surface patches and interfaces

Internal interfaces and molecular patches

Interfaces are deduced from the contacts of atoms listed in the interface file of the DIP (see the flow-chart

in Figure 1). Starting from the atomic coordinates of the PDB, the following procedure was used to find interacting sets of atoms. (1) The protein was dissected into SSEs. (2) For each protein, the distances between all pairs of atoms including hetero-atoms were calculated, excluding distances between atoms in the same amino acid residue or hetero-compound, or in residues adjacent in sequence. (3) An interface built by two neighboring MSPs was defined as follows: any atom of one particular secondary structure has at least one partner atom located on the other structural element, with a distance between the vdW spheres smaller than the chosen cut-off value. All atoms fulfilling the conditions of one particular SSE are collected in a particular MSP.

External molecular surface patches

For the external surface and the surface of larger holes in protein molecules, the position of neighboring atoms, e.g. the water molecules, are mostly unknown. The distance of protein atoms from a continuous solvent is defined using the molecular surface according to Connolly (1983). This exterior surface is generated by the closest possible vdW surfaces of virtual solvent atoms and can be used as a neighbor for superficial SSEs (see Figure 4). The distance from an atom to the solvent is defined as the minimal distance between the vdW and Connolly surfaces.

To verify that the different approaches for interior and exterior patches give similar results, we stripped largely inaccessible SSEs from their protein environment (e.g. the helix 161-177 from aldolase; PDB code 1ALD) and the cut-off dependence of atomic contacts was compared. Three-quarters of the atoms are found to be in contact at equal cut-off values, some at 0.5 Å lower for solvent contacts because of the assumed perfect atomic packing of the solvent around the protein (mean difference of cut-off values 0.11 Å, standard deviation 0.33 Å).

Evaluation of propensities of amino acids in SSEs and MSPs

As an example for the calculation of the propensities (as given in Table 4), the estimation of the value for glutamate in helix-extended patches ${}^{\mathrm{Glu}}\langle P\rangle_{\mathrm{HE}} = 0.75$ is explained and compared with the general helix propensity of glutamate ${}^{\mathrm{Glu}}\langle P\rangle_{\mathrm{H}} = 1.47$:

$$ {}^{\mathrm{Glu}}\langle P\rangle_{\mathrm{H}} = \frac{{}^{\mathrm{Glu}}\langle N\rangle_{\mathrm{H}}}{{}^{\mathrm{Glu}}\langle N\rangle}\frac{\langle N\rangle}{\langle N\rangle_{\mathrm{H}}} $$

$$ {}^{\mathrm{Glu}}\langle P\rangle_{\mathrm{HE}} = \frac{{}^{\mathrm{Glu}}\langle N\rangle_{\mathrm{HE}}}{{}^{\mathrm{Glu}}\langle N\rangle_{\mathrm{H}}}\frac{\langle N\rangle_{\mathrm{H}}}{\langle N\rangle_{\mathrm{HE}}} $$

with ${}^{\mathrm{Glu}}\langle N\rangle_{\mathrm{H}}$ the number of Glu residues in helices; $\langle N\rangle_{\mathrm{H}}$ the total number of residues in helices (24,694); $\langle N\rangle$ the total number of amino acid residues in the DIP (87,097); ${}^{\mathrm{Glu}}\langle N\rangle_{\mathrm{HE}}$ the number of Glu residues in helix-extended patches; $\langle N\rangle_{\mathrm{HE}}$ the total number of residues in helix-extended patches; and ${}^{\mathrm{Glu}}\langle N\rangle$ the number of Glu residues in the DIP.

Residues in patches are counted if at least one of their atoms is contained in the patch.

The computation of atomic packing

For this purpose, the volume occupied by a protein is divided into volumes related to each individual atom (Richards, 1979). Packing density is defined as the relation between the vdW volume of a given atom and the solvent-excluded volume associated with the atom. These volumes are computed according to the precise method of Goede *et al*. (1997).

The query system

The search tool was implemented with Delphi (object orientated Pascal with interface to database programming). Data fluxes are illustrated in Figure 1. The following types of parameters can be subjected to inquiries: length, depth, width, number of atoms, type of interface. Restrictions resulting from these conditions reduce the search space for similarity screenings. The following alignment procedure is outlined separately in the next section. All results are automatically evaluated according to atomic composition or local packing densities.

Automatic procedure for sequence-independent superposition

Because the definition of the MSPs exclusively considers spatial atomic neighborhood, a sequence-independent algorithm for their superposition was required for comparison of MSPs. Only a brief overview is given here, and further details will be published elsewhere. In a first step, the centers of mass of the patches are superimposed, followed by a rotation of one MSP such that the major directions (largest expansions) coincide. This normalization is used in a further step to determine the pairs of atoms between the two patches. Finally, the resulting superposition is expanded for neighboring atoms.

Tests concerning stability of the superposition procedure were carried out for patches consisting of 30 atoms, distributed in a typical parallelepiped with the lengths of edges 4, 8 and 12 Å. The alignment remained unaffected if up to seven atoms were randomly added to one of the patches within a sphere with a diameter of 14 Å.

Without optimization of the speed in this version, an alignment of two patches (30 atoms) takes about 0.1 second on a PC. A typical comparison between two (homologous) proteins with our comparison algorithm, which was tailored for different purposes, needs less than ten seconds on an IBM PC to find the best alignment.

References

Abagyan, R. & Maiorov, V. N. (1992). An automatic search for similar spatial arrangements of alpha-helices and beta-strands in globular proteins. *J. Biomol. Struct. Dynam.* **6**, 1045–1060.

Alesker, V., Nussinov, R. & Wolfson, H. J. (1996). Detection of non-topological motifs in protein structures. *Protein Eng.* **9**, 1103–1119.

Alexandrov, N. N. (1996). SARFing the PDB. *Protein Eng.* **9**, 727–732.

Alexandrov, N. N. & Go, N. (1994). Biological meaning, statistical significance, and classification of local spatial similarities in non-homologous proteins. *Protein Sci.* **3**, 866–875.

Alexandrov, N. N., Takahashi, K. & Go, N. (1992). Common spatial arrangements of backbone fragments in homologous and non-homologous proteins. *J. Mol. Biol.* **225**, 5–9.

Argos, P. (1988). An investigation of protein subunit and domain interfaces. *Protein Eng.* **2**, 101–113.

Bernstein, F., Koetzle, T., Williams, G., Meyer, E., Jr, Brice, M., Rodgers, J., Kennard, O., Shimanuchi, T. & Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). CHARMM: a program for macromolecular energy minimization and dynamics calculation. *J. Comput. Chem.* **4**, 187–217.

Brünger, A. T., Kuriyan, J. & Karplus, M. (1987). Crystallographic *R*-factor refinement by molecular dynamics. *Science,* **235**, 458–460.

Bures, M. G., Martin, Y. C. & Willet, P. (1994). Searching techniques for databases of three-dimensional chemical structures. *Topics Stereochem.* **21**, 467–511.

Chothia, C. (1975). Structural invariants in protein folding. *Nature,* **254**, 304–308.

Colloc'h, N., Etchebest, C., Thoreau, E., Henrissat, B. & Mornon, J.-P. (1993). Comparison of three algorithms for the assignment of secondary structure in proteins: the advantage of a consensus assignment. *Protein Eng.* **6**, 377–382.

Connolly, M. L. (1983). Solvent-accessible surfaces of proteins and nucleic acids. *Science,* **221**, 709–713.

Connolly, M. L. (1985). Computation of molecular volume. *J. Am. Chem. Soc.* **107**, 1118–1124.

Connolly, M. L. (1986). Shape complementarity at the hemoglobin a_1b_1 subunit interface. *Biopolymers,* **25**, 1229–1247.

Doolittle, R. F. & Bork, P. (1993). Evolutionarily mobile modules in proteins. *Sci. Am.* **269**, 50–56.

Eisenhaber, F. & Argos, P. (1993). Improved strategy in analytic surface calculation for molecular systems: handling of singularities and computational efficiency. *J. Comput. Chem.* **14**, 1272–1280.

Eisenhaber, F. & Argos, P. (1996). Hydrophobic regions on protein surfaces: definition based on hydration shell structure and a quick method for their computation. *Protein Eng.* **9**, 1121–1133.

Eisenhaber, F., Lijnzaad, P., Argos, P., Sander, C. & Scharf, M. (1995). The double cubic lattice method: efficient approaches to numerical integration of surface area and volume and for generating dot surfaces of molecular assemblies. *J. Comput. Chem.* **16**, 273–284.

Fasman, K. H., Cuticchia, A. J. & Kingsbury, D. T. (1994). The GDB human genome data base anno 1994. *Nucl. Acids Res.* **22**, 3462–3469.

Fischer, D., Tsai, C. J., Nussinov, R. & Wolfson, H. (1995). A 3D sequence-independent representation of the protein data bank. *Protein Eng.* **8**, 981–997.

Gerstein, M., Tsai, J. & Levitt, M. (1995). The volume of atoms on the protein surface: Calculated from simulation, using Voronoi polyhedra. *J. Mol. Biol.* **249**, 955–966.

Goede, A., Preißner, R. & Frömmel, C. (1997). Voronoi cell–a new method for allocation of space amongst atoms: avoidance of errors at volume and density calculations. *J. Comput. Chem.* **18**, 1113–1123.

Good, A. C., Ewing, T. J. S., Gschwend, D. A. & Kuntz, I. D. (1995). New molecular shape descriptors: application in database screening. *J. Comput. Aided Mol. Des.* **9**, 1–12.

Grindley, H. M., Artymiuk, P. J., Rice, D. W. & Willet, P. (1993). Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *J. Mol. Biol.* **229**, 707–721.

Hobohm, U. & Sander, C. (1994). Enlarged representative set of protein structures. *Protein Sci.* **3**, 522–24.

Holm, L., Ouzounis, C., Sander, C., Tuparev, G. & Vriend, G. (1992). A database of protein structure families with common folding motifs. *Protein Sci.* **1**, 1691–1698.

Hubbard, S. J. & Argos, P. (1995). Evidence on close packing and cavities in proteins. *Curr. Opin. Biotechnol.* **6**, 375–381.

Hubbard, S. J., Gross, K. H. & Argos, P. (1994). Intramolecular cavities in globular proteins. *Protein Eng.* **7**, 613–626.

Huysmans, M., Richelle, J. & Wodak, S. J. (1991). SESAM: a relational database for structure and sequence of macromolecules. *Proteins: Struct. Funct. Genet.* **11**, 59–76.

Janin, J. & Chothia, C. (1990). The structure of protein-protein recognition sites. *J. Biol. Chem.* **265**, 16027–16030.

Jones, S. & Thornton, J. M. (1995). Protein-protein interactions: a review of protein dimer structures. *Prog. Biophys. Mol. Biol.* **63**, 31–65.

Jones, S. & Thornton, J. M. (1996). Principles of protein-protein interactions. *Proc. Natl Acad. Sci. USA,* **93**, 13–20.

Kabsch, W. & Sander, C. (1983). Dictionary of secondary structures in proteins. *Biopolymers,* **22**, 2577–2637.

Kuhl, F. S., Crippen, G. M. & Friesen, D. K. (1984). A combinatorial algorithm for calculating ligand binding. *J. Comput. Chem.* **5**, 24–34.

Lawrence, C., Auger, I. & Mannella, C. (1987). Distribution of accessible surfaces of amino acids in globular proteins. *Proteins: Struct. Funct. Genet.* **2**, 153–161.

Lessel, U. & Schomburg, D. (1994). Similarities between protein 3D-structures. *Protein Eng.* **7**, 1175–1187.

Miller, S. (1989). The structure of interfaces between subunits of dimeric and tetrameric proteins. *Protein Eng.* **3**, 77–83.

Mizuguchi, K. & Go, N. (1995). Comparison of spatial arrangements of secondary structural elements in proteins. *Protein Eng.* **8**, 353–362.

Orengo, C. A., Flores, T. P., Taylor, W. R. & Thornton, J. M. (1993). Identification and classification of protein fold families. *Protein Eng.* **6**, 485–500.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). CATH–a hierarchic classification of protein domain structures. *Structure,* **5**, 1093–1108.

Padlan, E. A. (1990). On the nature of antibody combining sites: unusual structural features that may confer on these sites an enhanced capacity for binding ligands. *Proteins: Struct. Funct. Genet.* **7**, 112–124.

Pascarella, S. & Argos, P. (1992). A data bank merging related protein structures and sequences. *Protein Eng.* **5**, 121–137.

Peng, Z. Y., Wu, L. C., Schulman, B. A. & Kim, P. S. (1995). Does the molten globule have a native-like

tertiary fold? *Philos. Trans. Roy. Soc. ser. B*, **348**, 43–47.
Peters, K. P., Fauck, J. & Frömmel, C. (1996). The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *J. Mol. Biol.* **256**, 201–213.
Preißner, R., Goede, A. & Frömmel, C. (1997). Inverse sequence similarity in proteins and its relation to the three-dimensional fold. *FEBS Letters*, **414**, 425–429.
Richards, F. M. (1974). The interpretation of protein structures: total volume, group volume distributions and packing density. *J. Mol. Biol.* **82**, 1–14.
Richards, F. M. (1977). Areas, volumes, packing and protein structure. *Annu. Rev. Biophys. Bioeng.* **6**, 151–176.
Richards, F. M. (1979). Packing defects, cavities, volume fluctuations, and access to the interior of proteins including some general comments on surface area and protein structure. *Carlsberg Res. Commun.* **44**, 47–63.
Richmond, T. J. (1984). Solvent accessible surface area and excluded volume in proteins. *J. Mol. Biol.* **178**, 63–89.
Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Struct. Funct. Genet.* **9**, 56–68.
Stouten, P. F. W., Frömmel, C., Nakamura, H. & Sander, C. (1993). An effective solvation term based an atomic occupancies for use in protein simulations. *Mol. Simul.* **10**, 97–120.
Swindells, M. B., MacArthur, M. W. & Thornton, J. M. (1995). Intrinsic phi, psi propensities of amino acids, derived from the coil regions of known structures. *Nature Struct. Biol.* **2**, 596–603.
Thornton, J. M., McArthur, M. W., Smith, D. K., Gardner, S. P., Hutchinson, E. G., Morris, A. L. & Sibanda, B. L. (1990). Analysis of errors found in protein structure coordinates in the Brookhaven data bank. In *Accuracy and Reliability of Macromolecular Crystal Structures* (Henrick, K., Moss, D. S. & Tickle, I. J., eds), pp. 39–52, Science and Engineering Research Council, Daresbury Laboratory, Warrington, UK.
Tsai, C. J., Lin, S. L., Wolfson, H. J. & Nussinov, R. (1996). A dataset of protein-protein interfaces generated with a sequence-order-independent comparison technique. *J. Mol. Biol.* **260**, 604–620.
Vriend, G. & Sander, C. (1991). Detection of common three-dimensional substructures in proteins. *Proteins: Struct. Funct. Genet.* **11**, 52–58.
Williams, M. A., Goodfellow, J. M. & Thornton, J. M. (1994). Buried waters and internal cavities in monomeric proteins. *Protein Sci.* **3**, 1224–1235.
Wu, L. C., Grandori, R. & Carey, J. (1994). Autonomous subdomains in protein folding. *Protein Sci.* **3**, 369–371.

Appendix

```
HEADER    HYDROLASE(ENDORIBONUCLEASE)              25-SEP-91   9RNT
COMPND    RIBONUCLEASE T1 (E.C.3.1.27.3) COMPLEX WITH CA 2+
SOURCE    (ASPERGILLUS ORYZAE) RECOMBINANT FORM EXPRESSED IN
SOURCE   2 (ESCHERICHIA COLI)
AUTHOR    J.MARTINEZ-OYANEDEL,U.HEINEMANN,W.SAENGER
RESOLT    RESOLUTION. 1.5 ANGSTROMS.
METHOD    secondary structure analysis by DSSP
LENGTH     104 AA      779 ATOMS   783 ATOMS total
NRELEM     17 (number of elements)
NRHETE      4 (number of elements of HETATM)
NRHOLE      0 (number of cavities)
NRHOL2      3 (number of cavities without consideration of hetero-atoms)
OVERVW    C0001 E0004 C0007 E0009 C0012 H0013 C0029 E0040 C0043 E0056 C0061
OVERVW    E0076 C0082 E0086 C0092 E0101 C0103
OVERV2    O0780 W0781 W0782 W0783
SS   AA-No AA-ID Atom Atom-No   X      Y      Z   B-val  Vol1  Vol2 n SS1 ctoff  SS2 ctoff
C0001    1   ALA  N       1   6.504 -5.290 16.584 11.48 13.44 44.30 2  0    0.0   4    2.5
C0001    1   ALA  CA      2   5.998 -4.687 15.275 13.26 11.58 17.77 2  0    0.0   4    2.0
```

Figure A1. The Figure shows the start of an interface file. The content is: HEADER, protein class; COMPND, protein name; SOURCE, organism; AUTHOR, authors; RESOLT, resolution; METHOD, method for secondary structure analysis; LENGTH, number of amino acid residues, protein atoms, atoms (including hetero-complexes and inner water molecules); NRELEM, number of secondary structural elements; NRHETE, number of clusters of hetero-atoms; NRHOLE, number of cavities; NRHOL2, number of cavities excluding hetero-atoms; OVERVW, overview of the secondary structural elements (C, coil; E, extended; H, helix); the number is that of the starting amino acid residue of this element; OVERV2, overview of the hetero-clusters; O, other atoms; W, inner water molecules; SS, AA column headings, for an explanation see the text below; C0001, one line for any atom of the protein: the columns describing the contacts mean: backbone atoms N and C^{α} of coil 1 (coil starting with amino acid residue number one: C0001) are in contact with structure 0 (outside) at cut-off 0.0 Å and with secondary structure 4 (E0009) at cut-off 2.0 Å and cut-off 2.5 Å, respectively. The column headings are: SS, secondary structure containing this atom; AA-No, number of the amino acid residue (according to the PDB) containing this atom; AA-ID, three-letter code of the amino acid containing this atom; Atom, type of atom (according to the PDB); Atom-No, consecutive number of the atom; X, Y, Z, B-val, coordinates and *B*-value of the atom; Vol1 Vol2, vdW, volume and difference from solvent-excluded volume; n, number of contacts for this atom; SS1 ctoff, ID number for the first neighboring secondary structure and corresponding cut-off; SS2 ctoff, the same for the second contact.

Edited by R. Huber

(Received 15 December 1997; received in revised form 9 April 1998; accepted 9 April 1998)