Hydrophobic patches on protein surfaces

Hydrofobe gebieden op eiwitoppervlakken

(Met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht
op gezag van de rector magnificus, prof. dr. W.H. Gispen, ingevolge
het besluit van het college voor promoties in het openbaar te
verdedigen op

donderdag 26 april 2007 des middags te 4.15 uur.

door

Philippus Lijnzaad
geboren op 1 februari 1965 te Rotterdam

Promotoren: prof. dr. F.C.P. Holstege
prof. dr. J. Heringa

What a beautiful world this will be
What a glorious time to be free

Donald Fagan, I.G.Y.

Voor Catrien
en Dorien
en Maaike
en Fleur

Cover: Porcine phospholipase A_2 has a large, persistent hydrophobic patch that is probably functional. It consists of Leu 58 and Phe 94, and is remote from the active site and the phospholipid binding face of the protein, located behind helices C and E. The region around the novel patch is shown by its solvent accesible surface; the remainder is represented as a cartoon. (Image by Raster3D, Merrit & Bacon (1997), Meth. Enz. **277**:505-524)

Layout: By the author, using the LaTeX 2_ε Documentation System on a PC running Ubuntu GNU/Linux.
Print: Ponsen & Looijen, Wageningen
ISBN: 978-90-393-4507-8

CONTENTS

1

GENERAL INTRODUCTION

CODES OF LIFE

The information flow in cellular processes can be summarized loosely as follows:

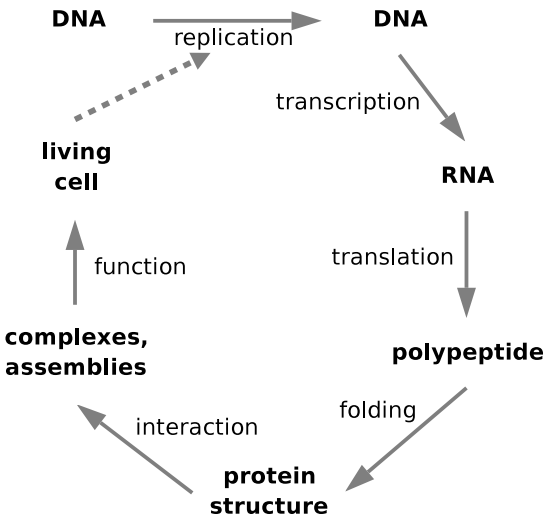

Figure 1: *Summary of Molecular Biology*

This cartoon is an elaboration of molecular biology's central dogma "DNA makes RNA makes protein". The steps *replication*, *transcription* and *translation* are well understood: we can predict the entities in the next stage, given those in the previous one. Our ability to predict stems from knowing the relevant *code*. The DNA-to-DNA code is simply the base-pairing rule, and the code to make RNA from DNA is similar. The code linking RNA sequence to protein sequence is the genetic code, but we do not know the code for the the subsequent steps: *folding*, *interaction* and *function*.

It has become clear that there is no simple "folding code" that allows us to predict the protein structure from the amino acid sequence alone. This is unfortunate, as the panoply of proteins is responsible for executing most of the processes in a living cell. A protein achieves its function through its three-dimensional, folded structure, and the elucidation of it through X-ray crystallography or NMR methods invariably provides rich insights into its workings.

Although there is no simple folding code, many inroads have been made in recent years. This progress is only partly due to a better understanding of the

physics underlying the folding process; much of the physics was reasonably clear by the mid-70's.[1] The laws of physics and chemistry do, of course, constitute a kind of universal "code", but it is not a very practical one, because of the huge number of degrees of freedom of a polypeptide. This makes strictly *ab initio* prediction of protein structure using solely physical methods infeasible, in spite of the enormous gains in computer power over the last 30 years.

The progress that *has* been booked in predicting structure from sequence is due largely to homology methods, which detect and exploit even faint homologies with sequences of known structure.[2] These prediction endeavours have benefited greatly from the enormous increase in the number of protein structures determined experimentally. An interesting phenomenon in this respect is the organization, since 1996, of the Critical Assessment of Structure Prediction (CASP) competition.[3] In it, the participating research groups are given sequences of proteins that crystallography groups have solved, but not yet published. The participants then pit their programs and expertise against each other to predict the protein structure as accurately as possible. The competition consists of various categories (levels of homology; degree of human intervention, folding, prediction of domain boundaries, *etc*.), and it provides a level playing field to gauge the progress in our ability to crack the "folding code". Ref. 3 suggests that in another ten years, homology-based protein models should be of atomic resolution.

A problem slightly less intractable than *ab initio* folding is the next step in Fig. 1: *interaction*. Physical interactions between proteins, and between proteins and smaller compounds sometimes *are* amenable to Molecular Mechanics approaches. In such an approach, the energy of a complete (near-)atomic model of both the compounds (proteins, ligands, solvent, ions) and their internal and external interactions (the so-called *force field*) is minimized *in silico* in order to obtained the most stable conformation or association. During such *protein docking* attempts, the backbone of the protein is typically held rigid, substantially reducing the number of degrees of freedom. With these simplifications and given enough computer power, the physics "code" sometimes *is* employed successfully in predicting the three-dimensional details of protein-protein or protein-ligand complexes. The protein docking field is distinct from the protein modelling field, but it has a similar competition: CAPRI (Critical Assessment of PRedicted Interactions).[4]

Predicting interactions between a protein molecule and DNA is much more difficult, because the large flexibility of DNA is essential in the recognition by DNA-binding proteins. This makes rigid body approaches unfruitful, but progress is being made.[5]

STRUCTURAL BIOLOGY

The last step in our diagram, *function*, is, in fact, the aim of structural biology: understanding the *functional code*, that is, explaining how a protein's function arises

from its structure*. Or, put even stronger, to predict function from structure. Although we can currently *predict* function only using homology methods, it is not inconceivable that function could be suggested by structure alone. For instance, finding a surface cleft predicted to be able to bind ATP would certainly be suggestive of ATPase activity as one of the protein's functions. This could be a useful adjunct to genome-scale prediction of protein structure using *ab initio* or homology methods, and such an approach could have a place in the context of *structural genomics* projects.[6] However, the functional code – if that is at all an appropriate metaphor – is not a simple one. Often, simply binding or recognizing another protein, a ligand, a substrate, or DNA, explains function adequately. But even if much about a protein (or gene) is known from biochemical and genetical studies, it is not always obvious, at first sight, how a protein structure (or protein complex) achieves its function. For instance, little is known about the actual mechanism of electron transfer in redox proteins, even though extensive structural and biological information have long been available.

A number of common themes in structure-function relationships is nonetheless well-understood. *E.g.*, the main function of structural proteins such as actin and collagen is to provide relatively homogeneous, inert molecular scaffolding. Another possibility is providing *volume*, as illustrated by the crystallins. They provide homogeneous optical density, in bulk, to the eye lens, with exceptional stability. Motor proteins such as flagellin and myosin are essentially nanometer-scale machines. The same can be said of membrane channel-forming proteins such as the nuclear pore complex, which exert their function partly sterically as a structural protein, and partly as a molecular machine, when selectively opening or closing. Many regulatory proteins act by blocking access to an active site, or by physically impeding progress of another cellular entity. Proteins such as myoglobins or ferredoxins can be thought of as "merely" a binding platform for non-protein compounds that do the real work (here: heme to bind molecular oxygen, and an iron-sulfur cluster to carry electrons, respectively). Maybe the most well-studied class of structure-function relationships is that of catalysis by enzymes. Here, the function can be divided in two parts: that of *molecular recognition* of the substrate(s), and that of the catalysis *per se*. The latter is achieved by stabilizing the transition state of the chemical reaction catalyzed.

This list of themes is far from exhaustive, but the unifying abstraction that emerges is that of the duo *recognition* and *binding* (corresponding to the interaction "code" depicted in Fig. 1). They could be called *meta-functions*, where recognition confers specificity, and binding the ability to assemble greater functionality from smaller parts.

Much attention is often given to the protein fold; that is, to the overall architecture of the protein structure, its secondary structure elements and their interrelations, *etc*. While this is clearly important from the point of view of protein stability and evolution, it is at the protein surface that recognition and binding take place.

*It is misleading to speak of *the* function of a protein; it often has, or may have, several.

HYDROPHOBICITY

An important role in molecular recognition and association is played by hydrophobicity; it is also one of the main players in protein folding, as will be elaborated on below.

The association of protein subunits into larger complexes is an important phenomenon in structural biology. Usually, a distinction is made between strong complexes and weak complexes. The strong, so-called *obligate* complexes have dissociation constants in the nanomolar range, and their monomers are deemed not to be able to occur independently. In contrast, the weaker, non-obligate complexes have dissociation constants in the micromolar range, and are more transient. Subunit assocation in obligate oligomeric protein complexes resembles partly molecular recognition, and partly protein folding. The subunits are slightly less stable than typical monomeric proteins, and the association delivers additional stabilization. The interface between the subunits looks somewhat like the hydrophobic core of monomeric proteins, although it is less hydrophobic.[7, 8] Chothia and Janin[9] showed that in subunit association of obligate complexes, hydrophobicity is the dominant driving force. The same is true for the binding of small, mostly hydrophobic compounds. The situation is less clear for non-obligate complexes; here, hydrophobicity does play a role, but a less prominent one.[10, 11]

What is clear is that surface hydrophobicity is crucial in functional respect. This is also supported by functional genomics approaches: Deeds *et al.*[12] observe a strong correlation between the hydrophobicity of a protein and its number of interacting partners in protein-protein interaction networks.

Although nature employs surface hydrophobicity extensively for recognition and binding, it also involves a cost: exposing hydrophobic surface area to the solvent is destabilizing. The reason for this is the *hydrophobic effect*. It is the tendency of polar solutions, such as water, to exclude oily (=hydrophobic) solutes. It arises *not* because apolar solutes are intrinsically repelled by water; in fact, the van der Waals interactions between apolar compounds and water are generally favourable. It is instead due to the strong hydrogen-bonding capacity of water. Water molecules in liquid water are tetrahedrically coordinated, and the large majority is involved in four hydrogen bonds per water molecule. This three-dimensional mesh of strong hydrogen bonds is responsible for many of the anomalous properties of water (for a popular account, see Ref. 13, Chapter 6). The mesh is very dynamic, as there is enough freedom for a water molecule to reorientate without breaking all its hydrogen bonds. This becomes different when a water molecule is in the direct vicinity of a solute with which it cannot hydrogen-bond. In this situation, most water molecules will still be able to fulfill their "need" for four hydrogen bonds, with other water molecules. In fact, these hydrogen bonds are slightly stronger than those in bulk water, and are sometimes referred to as "iceberg" water.[14] However, this slight gain in stability is more than off-set by a large entropic cost, because in the new situation, water molecules do not have full freedom to reorient without breaking most of their hydrogen bonds. In other words, the epi-

thet "hydrophobic" for apolar substances is a misnomer; it would be more accurate (although not very helpful) to call water *lipophobic*.

The entropic cost incurred by hydrophobic solvation is proportional to the number of water molecules being in contact with apolar surface area. Since it dominates the free energy of solvation, this ΔG_{solv} must be expected to vary linearly with the extent of the surface area. This is indeed found to be the case.[15, 16] The exact constant of proportionality is called the specific hydrophobic surface solvation free energy, or briefly σ_{solv}. It represents the energy, per mol, needed to expose one square Ångstrom* of hydrophobic surface to water. The experimental systems used to measure or infer this quantity vary widely; most involve the transfer of solutes from apolar solvent to water, but many other ways have been devised. The resulting values for σ_{solv} are not consistent, with values ranging between +8 and +73 cal/Å^2 mol** (for references, see Refs. 17 and 18). Hydrophobic interaction is semi-quantitatively understood as arising from water structure, and although progress has been booked (see *e.g.* Ref. 19), our understanding is still incomplete (for a recent review, see Ref. 20).

PROTEIN FOLDING

Water's propensity to force apolar groups together is considered one of the main driving forces behind the folding of water-soluble proteins†. Approximately one third of the 20 amino acid types is apolar, and these amino acids make up most of the strongly hydrophobic core of most protein structures, with the more polar groups constituting the surface of the protein. This *hydrophobic inside, hydrophilic outside* picture was initially compared with the formation of *micelles*, the clusters into which fatty acids and other amphipathic molecules assemble when dissolving in water. However, this picture is too simple. Compared to micelles, protein structures are almost crystalline, with a highly structured, densely packed interior and low mobilities for internal groups. Also, protein structures have other forces stabilizing them. Prime among them, of course, are the internal hydrogen bonds that are responsible for the occurrence of regular secondary structure such as α-helices and β-sheets, as well as the various types of turns. Another important presence in the core of protein structures, also not found in micelles, is that of so-called *internal waters*. They are relatively fixed water molecules with long residence times that are an integral part of the interior hydrogen-bonding network of the protein‡.

Although it is now generally accepted that hydrophobicity is the dominant force driving protein folding,[23–28] this consensus took time to develop. Privalov and others concluded that the hydrophobic effect *destabilizes* the native protein structures

*One Ångstrom is 10^{-10} m.

**1 cal = 4.184 J.

†Transmembrane proteins are ignored in this work.

‡For this reason, internal water is sometimes called "the 21st amino acid".[21, 22]

(discussed extensively in Ref. 29, Chapter 7). The issue seems to have been resolved by proper statistical-mechanical treatment of the curvature of the hydrophobic surface and of drying effects.[30–32]

A less controversial (and less testable) issue in protein folding is that stable protein structures must be in a global free energy minimum.[33, 34] This is the so-called *thermodynamic hypothesis*, and forms the basis for a plethora of statistical "potentials" used to assess and even predict protein structure. This is very popular; Shen and Šali[35] cite 42 different, so-called *knowledge-based* potentials, all derived from statistical analysis of protein structures in the Protein Data Base*.

However, this free energy minimum must also be achievable in finite time.[36] The latter demand imposes constraints on protein composition (see *e.g.* Refs. 37 and 38), but also on the primary sequence, as not just any protein sequence is likely to fold into a stable structure.[39–43] Current thinking is that evolution has helped make the so-called *energy landscape* more funnel-like, allowing smooth folding to the native structure.[40, 43–46] The same is thought to hold for protein-protein interactions.[47, 48]

Not all of a protein structure is always ordered, or even completely folded, in the native, functional state. It has long been known that surface groups are mobile, and this is especially true for turns and side chains; frequently, they can not even be resolved in the electron density maps obtained from X-ray crystallography. A more interesting case is that of the so-called *intrinsically disordered proteins*.[49] Their (partial) lack of structure is observed to be functionally important in many areas of cellular biology, primarily by enabling easy posttranslational modifications such as phosphorylation.

Disorder more frequently implies dysfunction, as is the case with protein denaturation, which can be subdivided into aggregation, (local) unfolding and misfolding. Denaturation can occur under a variety of circumstances. The most familiar one is heat, but molecular crowding, low temperature, high pressure or aberrant solvent composition are other important causes of unfolding. Cells cope with this through a class of proteins called molecular chaperones. They recognize and stabilize (partially) unfolded proteins, an important special case being the nascent polypeptide chain during translation.

The mode of action of the GroEL/GroES chaperonin system is known in some detail,[50, 51] and is remarkable. The hydrophobic regions of folding and misfolded proteins bind the hydrophobic walls of the GroEL complex, which forms a large cage. Upon closing by the GroES "cap", a massive domain shift causes the walls of the cage to become hydrophilic. This provides a stabilizing environment ("Anfinsen cage"), where the protein can fold in isolation, without the possibility of aggregation, as it were at infinite dilution. After a fixed period of approximately 10 s, the complex reopens, and the protein is released; if folding was not completed, another round ensues. Chaperones assist folding largely non-specifically, and do not influence the exact structure of the folded protein, which is fully determined by

*And these were all publications of 1990 and later.

the amino acid sequence.

A large number of pathologies is caused by misfolding and aggregation, and proteins likely to misfold appear to be overrepresented amongst proteins that are involved in disease.[46] These disorders are collectively known as amyloidoses, and can be both hereditary and acquired. They are caused by the disruption of cellular function as a result of the formation of large protein deposits called amyloid plaques. Well-known examples are Type II diabetes, and Alzheimer's, Parkinson's, Huntington's and Creuzfeld-Jacob disease. The underlying common theme is the formation of large fibrillar protein aggregates, consisting of extensive intermolecular β-sheets;[52] for a review of the biophysics, see Ref. 53. In the initial and potentially most cytotoxic stages of the formation of the aggregate however, intra- or intermolecular hydrophobic interactions and/or aberrant association with membranes appear to be involved.[52–56]

PROTEIN SURFACES

Clearly, hydrophobicity is essential in the physics of protein (mis)folding, and we can calculate the free energy cost of exposing hydrophobic surface to the solvent simply by multiplying its area with aforementioned σ_{solv}. A direct consequence of this proportionality is that the (de)solvation free energy of folding of a protein is also proportional to the length of the protein sequence. This has been confirmed experimentally for a wide range of soluble protein structures;[57] the slope of the regression line linking sequence length and solvation free energy is roughly 1 kcal per residue per mol. This fact is an important quantitative quality criterion when judging the accuracy of a predicted protein structure: if too little hydrophobic surface area of the initially extended polypeptide chains is buried upon (predicted) folding, the accuracy of the model is probably poor.[58, 59]

How, then, is the protein surface defined and measured? Lee and Richards have proposed the following terminology,[60] which is illustrated in Fig. 2. The *van der Waals* (or *atomic*) surface is the surface that is obtained when the atoms are represented as partly overlapping hard spheres; it is the surface most often used for representational purposes. It is typically rendered with molecular graphics software such as RasMol,[61] selecting the "CPK"[62] or "space-filling" display option. The van der Waals surface is less relevant from the point of view of solvent interaction, as its area does not measure the extent to which the solute can interact with the solvent. For instance, atom *A* in Fig. 2a lies in a small cleft, and is not in direct contact with water, represented by the probe located at position *p1*. So although atom *A* contributes to the van der Waals surface and would be visible using molecular graphics software, it is not *solvent accessible*, and is usually considered buried.

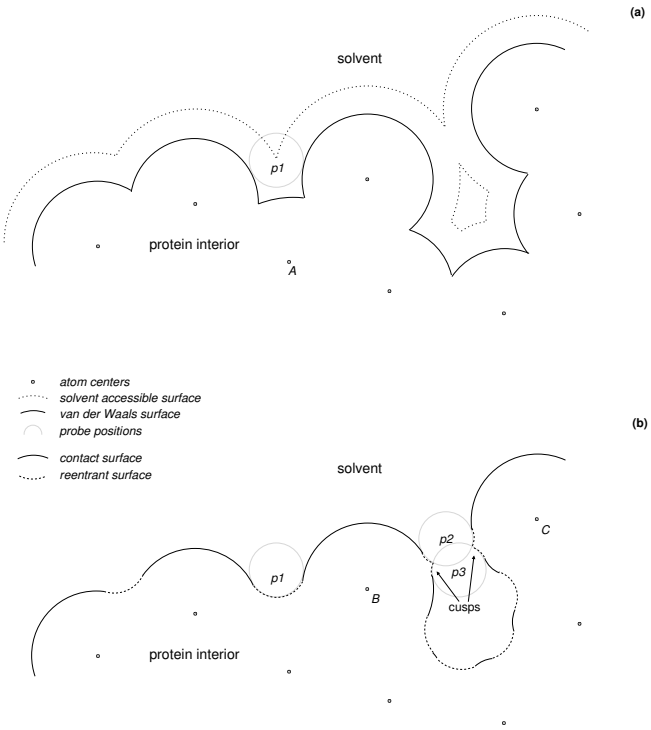

Figure 2: *Protein surface definitions. The sections through the protein are identical. For a detailed explanation, see text. (a) The van der Waals surface and solvent accessible surface. (b) The molecular surface.*

The surface that is more relevant in the study of protein folding and hydrophobicity is the so-called *solvent accessible surface*. It is a virtual surface, and is the boundary of the volume that is accessible to the center of a solvent molecule. It is obtained by representing the solvent molecule (usually water) as a sphere, traditionally called the *probe* (see Fig. 2a). The probe is rolled over the complete van der Waals surface. The surface traced out by the center of the probe is then the solvent accessible surface. In other words, the solvent accessible surface is identical to the van der Waals surface, if the radius of the probe (usually 1.4 Å) is added to that of all the atoms prior to the calculation, and this is precisely how it is done in practice.

There are a few more definitions in common use: the *molecular surface*, which

itself consists of *contact* and *reentrant* surface (see Fig. 2b). The molecular surface is sometimes called the Connolly surface, after Michael Connolly who pioneered its calculation.[63] It is the surface that results when "shrink wrapping" the atoms with an imaginary foil that is pressed into the clefts by the probe. Where the probe touches the atoms, this surface is called the contact surface; here, it coincides with the van der Waals surface. Where the probe cannot touch the atoms, the foil assumes the shape of the probe and is concave, bridging the contact surfaces of the neighbouring atoms. This occurs at the "joint" between two or three atoms, as for instance shown in Fig. 2b by the probe located in position *p1*. This concave part of the molecular surface is called the reentrant surface. This surface is somewhat problematic, in that it does not "belong" to one atom, and it may contain so-called *cusps*. They occur when the reentrant surface intersects itself, as shown in Fig. 2b. Here, the probe size and the distance between atoms *B* and *C* are such that the probe, "rolling" around an axis formed by the line *B*-*C* as indicated by positions *p2* and *p3*, leaves two cusps on atoms *B* and *C*. With the exception of the cusps, the molecular surface is smooth; this has made it attractive for visualization purposes.

There are analytical and numerical approaches to calculating the surface, both for display purposes and for obtaining the area, and both for the van der Waals or solvent accessible surface as well for the molecular surface. The analytical determination of the surfaces is fairly complicated. In the case of the van der Waals or solvent accessible surface, intersection circles are determined for each pair of neighbouring atoms. The intersection circles of triplets of neighbouring atoms are subsequently combined to determine the arcs that bound the surface belonging to each atom. Round-off errors have to be addressed, given the enormous number of arcs involved.[64] Calculating the molecular surface analytically is similar, but now one has to deal with toroidal segments (the reentrant surface between two atoms) and concave spherical triangles (the reentrant surface between a triplet of atoms).[63] Totrov and Abagyan[65] later improved on both the run-time behaviour and the treatment of cusps of the Connolly algorithm.

The numerical approximation of the solvent accessible surface (or van der Waals surface) is far simpler. It consists of placing a spherical point distribution on each atom, and determining which of the points is buried by any other atoms.[66] The remainder of the points is accessible, and can be used directly for visualization. The (weighted) sum of accessible points is a measure of the area per atom. The main task is to make the test for burial fast. As often, accuracy and speed are in opposition: using a spherical distribution with few points to sample the surface is fast, but the approximation is rough. Conversely, densely covering the atoms with points in order to improve the accuracy will require longer times testing for burial. In a naive implementation, the number of burial tests is proportional to the product of the point density and the square of the number of atoms, making it a costly operation. A numerical determination of the molecular surface is provided by the SIMS method,[67] which also removes cusps.

It is not entirely clear which of these surface definitions is the proper one to calculate the hydrophobic solvation free energy. Most often, the solvent accessible surface area is used, both when deriving σ_{solv} and when calculating the total hydrophobic solvation free energy of a particular protein structure. Some authors claim that the molecular surface, rather than the solvent accessible surface, is the appropriate concept to use.[30, 68] An important advantage of the solvent accessible surface in this respect is the ease with which it can be obtained numerically; this probably has contributed considerably to its dominance.

As explained before, it is misleading to think of hydrophobic interaction as an attraction between apolar groups. However, it is still useful to speak of hydrophobic interaction, or even hydrophobic force, even though this force arises as the combined result of water ordering. The hydrophobic force can even be measured with atomic force microscopy, and turns out to have a very long range of up to 200 Å, although details still engender controversy.[20]

The fact that the hydrophobicity appears to act as a force between apolar groups has lead to on-going attempts to integrate the presence of the solvent into the solvent in an implicit way. This would obviate the need for the explicit simulation of solvent, thus enabling longer simulation times. This is especially welcome in the context of theoretical protein folding studies. The currently maximum attainable simulation runs, representing hundreds of nanoseconds, are too short for protein folding, which in occurs on the microsecond timescale. Results are varied,[68–71] but continue to improve.[72–75]

A problem inherent in the molecular and van der Waals or solvent accessible surfaces is that there is no simple derivative of the area with respect to the atomic coordinates, be they cartesian or internal. The reason is simply the complexity of the non-local and non-smooth interactions that lead to the formation of the surface, yielding an ill-behaved system. This makes surface calculations largely irrelevant to molecular dynamics force fields. Molecular dynamics calculations now generally include thousands of explicit water molecules. They model the intricacies of the hydrophobic effect automatically, and have the additional advantage of treating the polarization and electrostatic shielding more faithfully. The lack of derivatives poses no problem to Monte Carlo methods, as they (in the strictest sense) only require the ability to evaluate the total energy. They are therefore free to use surface area terms in their force field, representing implicit solvation energies.[76]

For the sake of completeness, it should be mentioned that the Gauss theorem, linking surface and volume integrals, allows the calculation of molecular volumes from the surfaces obtained by any of the methods discussed.

HYDROPHOBIC PATCHES

It appears to be under-appreciated that the surface of most soluble proteins is rather hydrophobic. Most of their solvent accessible surface is taken up by carbon and sulphur atoms, yielding an area that is, on average, around 60% hydrophobic.

A consequence is that hydrophobic regions on the proteins surface are ill-defined, simply because most of the protein surface is one contiguous hydrophobic region; see Fig. 3.

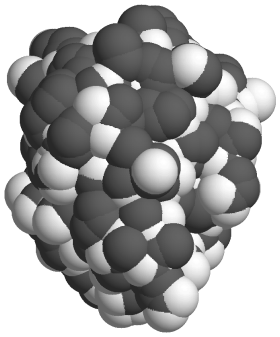

Figure 3: *The solvent accessible surface of a typical protein. White: hydrophilic atoms (oxygen and nitrogen); grey: hydrophobic atoms (carbon and sulfur). Hydrogens are not depicted.*

Not much research has been done regarding the local atomic details of hydrophobicity at the protein surface, whereas this would seem desirable from the foregoing discussion. In many cases, crystallographic researchers use an *ad hoc* definition of "hydrophobic patch" to describe their subjects of study; see for instance refs.[77–79] Similarly, the interface between multi-subunit protein complexes is often described in terms of hydrophobic patches.[80–83] In other words, the notion of a "hydrophobic patch" appears to be a useful one, but also one that is in need of a rigorous definition. This should be a useful addition to the range of protein structural analysis tools already available.

In many cases, the level of resolution at which the patches are (informally) defined is that of amino acid residues. The advantage of this is that it is simple, and can be easily generalized and studied in evolutionary terms. However, it is certainly possible and in fact observed that hydrophobic patches are composed of various atoms from different residues, not even all of them conventionally regarded as hydrophobic. This could have pharmaceutical implications, as exemplified by methotrexate, a folic acid analogon used in the treatment of cancer and autoimmune diseases. Its benzene ring appears to interact with the C-γ and C-δ of Lys32, when bound to *E. coli* dihydrofolate reductase (PDB code 1dds). This is supported by the low temperature factors of the lysine atoms.

The current work is at least partially inspired by the physical chemistry of proteins, and as a result, the atomic view formed the starting point for the definition of hydrophobic patches.

OVERVIEW OF THIS THESIS

The aim of this thesis is firstly to develop a rigorous definition of hydrophobic patches, as well as a method for their efficient automated detection. Secondly, the method is employed for a number of surveys of the protein surface in general. This is the primary focus of the work, and it is adopted in the belief that a better description and deeper understanding of the protein surface is important in judging functional aspects. Insights emanating from the overviews contribute to the more general understanding of protein structure, and should be useful in protein modelling, protein engineering, functional studies and drug design. Some cases of functionally interesting hydrophobic patches are encountered and discussed, strengthening the validity and utility of the method.

The basis for the implementation of the patch detection method is formed by a highly efficient numerical approximation of the solvent accessible surface called the Double Cubic Lattice Method (DCLM), which is described in Chapter 2. Chapter 3 describes the method (called QUILT) that defines and detects hydrophobic patches. It includes a measure to judge their likely functional significance, and is applied to a number of illustrative examples. QUILT is subsequently used to study the surface of monomeric proteins in general in Chapter 4, while Chapter 5 describes the trends regarding hydrophobic patches in interfaces of obligate protein complexes. In order to gain insight into the dynamic behaviour of patches, the method is extended and applied to Molecular Dynamics trajectories of three proteins; this is the subject of Chapter 6. Lastly, Chapter 7 is a general discussion of the results obtained, and addresses some of the implications and interpretation of the results, as well as pointers to future research.

BIBLIOGRAPHY

[1] L.R. Pratt and D. Chandler. Theory of the hydrophobic effect. *J. Chem. Phys.*, 67:3683–3704, 1977.

[2] O. Schueler-Furman, C. Wang, P. Bradley, K. Misura, and D. Baker. Progress in Modeling of Protein Structures and Interactions. *Science*, 310:638–642, 2005.

[3] A. Kryshtafovych, C. Venclovas, K. Fidelis, and J. Moult. Progress over the first decade of CASP experiments. *Proteins*, 61:225–236, 2005.

[4] R. Méndez, R. Leplae, M.F. Lensink, and S.J. Wodak. Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. *Proteins*, 60:150–169, 2005.

[5] M. van Dijk, A. D. J. van Dijk, V. Hsu, R. Boelens, and A.M.J.J. Bonvin. Information-driven protein-DNA docking using HADDOCK: it is a matter of flexibility. *Nucleic Acids Research*, 34:3317–3325, 2006.

[6] A. Rossi and A. Marti-Renom, M.A. Šali. Localization of binding sites in protein structures by optimization of a composite scoring function. *Prot. Sci.*, 15:2366–2380, 2006.

[7] J. Janin, S. Miller, and C. Chothia. Surface, subunit interfaces and interior of oligomeric proteins. *J. Mol. Biol.*, 204:155–164, 1988.

[8] N. Horton and M. Lewis. Calculation of the free energy of association for protein complexes. *Prot. Sci.*, 1:169–181, 1992.

[9] C. Chothia and J. Janin. Principles of protein-protein recognition. *Nature*, 256:705–708, 1975.

[10] I.M.A. Nooren and J.M. Thornton. Diversity of protein-protein interactions. *EMBO J.*, 22:3486–3492, 2003.

[11] J. Fernandez-Recio, M. Totrov, C. Skorodumov, and R. Abagyan. Optimal docking area: A new method for predicting protein-protein interaction sites. *Proteins*, 58:134–143, 2005.

[12] E.J. Deeds, O. Ashenberg, and E.I. Shakhnovich. A simple physical model for scaling in protein-protein interaction networks. *Proc. Natl. Acad. Sci.*, 103:311–316, 2006.

[13] P. Ball. H_2O. *A Biography of Water.* Phoenix/Orion Books Ltd, London, 2000.

[14] H.S. Frank and M.W. Evans. Free volume and entropy in in condensed systems. *J. Chem. Phys.*, 13:507–523, 1945.

[15] J.A. Reynolds, D.B. Gilbert, and C. Tanford. Empirical correlation between hydrophobic free energy and aqueous cavity surface area. *Proc. Natl. Acad. Sci.*, 71:2925–2927, 1974.

[16] S.C. Valvani, S.H. Yalkowsky, and G.L. Amidon. Solubility of nonelectrolytes in polar solvents. VI. Refinements in molecular surface area computations. *J. Phys. Chem.*, 71:829–835, 1976.

[17] A.H. Juffer, F. Eisenhaber, S. J. Hubbard, D. Walther, and P. Argos. Comparison of atomic solvation parametric sets: Applicability and limitations in protein folding and binding. *Prot. Sci.*, 4:2499–2509, 1995.

[18] F. Eisenhaber. Hydrophobic regions on protein surfaces. Derivation of the solvation energy from their area distribution in crystallographic protein structures. *Prot. Sci.*, 5:1676–1686, 1996.

[19] T.M. Raschke, J. Tsai, and M. Levitt. Quantification of the hydrophobic interaction by simulations of the aggregation of small hydrophobic solutes in water. *Proc. Natl. Acad. Sci.*, 98:5965–5969, 2001.

[20] E.E. Meyer, K.J. Rosenberg, and J. Israelachvili. Recent progress in understanding hydrophobic interactions. *Proc. Natl. Acad. Sci.*, 103:15739–15746, 2006.

[21] D. Russo, G. Hura, and T. Head-Gordon. Hydration Dynamics Near a Model Protein Surface. *Biophys. J.*, 86:1852–1862, 2004.

[22] G.A. Papoian, J. Ulander, M.P. Eastwood, Z. Luthey-Schulten, and P.G. Wolynes. Water in protein structure prediction. *Proc. Natl. Acad. Sci.*, 101:3352–3357, 2004.

[23] W. Kauzmann. Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.*, 14:1–63, 1959.

[24] A. Nicholls, K.A. Sharp, and Barry Honig. Protein folding and association: Insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins*, 11:281–296, 91.

[25] G. Casari and M.J. Sippl. Structure-derived hydrophobic potential. Hydrophobic potential derived from X-ray structures of globular proteins is able to identify native folds. *J. Mol. Biol.*, 224:725–732, 1992.

[26] C. Tanford. How protein chemists learned about the hydrophobic factor. *Prot. Sci.*, 6:1358–1366, 1997.

[27] H. Li, C. Tang, and N.S. Wingreen. Nature of Driving Force for Protein Folding: A Result From Analyzing the Statistical Potential. *Phys. Rev. Lett.*, 79:765–768, 1997.

[28] D. Chandler. Interfaces and the driving force of hydrophobic assembly. *Nature*, 437:640–647, 2005.

[29] T.E. Creighton. *Proteins. Structures and molecular properties. Second Edition*. Freeman, New York, 1993.

[30] K.A. Sharp, A. Nicholls, R.F. Fine, and B. Honig. Reconciling the magnitude of the microscopic and macroscopic hydrophobic effects. *Science*, 252:106–109, 1991.

[31] K. Lum and J.D. Chandler, D. Weeks. Hydrophobicity at small and large length scales. *J. Phys. Chem.*, B 103:4570 –4577, 1999.

[32] D.H. Huang and D. Chandler. Temperature and length scale dependence of hydrophobic effects and their possible implications for protein folding. *Proc. Natl. Acad. Sci.*, 97:8324–8327, 2000.

[33] C. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181:223–230, 1973.

[34] G. Govindarajan and R.A. Goldstein. On the thermodynamic hypothesis of protein folding. *Proc. Natl. Acad. Sci.*, 95:5545–5549, 1998.

[35] M. Shen and A. Šali. Statistical potential for assessment and prediction of protein structures. *Prot. Sci.*, 15:2507–2524, 2006.

[36] C. Levinthal. Are there pathways for protein folding? *J. Chim. Phys.*, 65:44–45, 1968.

[37] D.K. Klimov and D. Thirumalai. Factors Governing the Foldability of Proteins. *Proteins*, 26, 1997.

[38] B.G. Ma, J.X. Guo, and H.Y. Zhang. Direct correlation between proteins' folding rates and their amino acid compositions: An ab initio folding rate prediction. *Proteins*, 65:362–372, 2006.

[39] A. M. Gutin, V. I. Abkevich, and E. I. Shakhnovich. Chain Length Scaling of Protein Folding Time. *Phys. Rev. Lett.*, 77:5433–5436, 1996.

[40] C.-J. Tsai, S. Kumar, B. Ma, and R. Nussinov. Folding funnels, binding funnels, and protein function. *Prot. Sci.*, 8:1181–1190, 1999.

[41] I.N. Berezovsky, V.M. Kirzhner, A. Kirszhner, and E.N. Trifonov. Protein folding: Looping from Hydrophobic Nuclei. *Proteins*, 45:346–350, 2001.

[42] G. Dantas, D. Kuhlman, B. ăand Callender, M. Wong, and D. Baker. A Large Scale Test of Computational Protein Design: Folding and Stability of Nine Completely Redesigned Globular Proteins. *J. Mol. Biol.*, 332:449–460, 2003.

[43] J.N. Onuchic and P.G. Wolynes. Theory of protein folding. *Curr. Op. Struct. Biol.*, 14:70–75, 2004.

[44] K.A. Dill and H.S. Chan. From Levinthal to pathways to funnels. *Nat. Struc. Mol. Biol.*, 4:10–19, 1997.

[45] T. Head-Gordon and S. Brown. Minimalist models for protein folding and design. *Curr. Op. Struct. Biol.*, 13:160–167, 2003.

[46] P. Wong and D. Frishman. Fold Designability, Distribution, and Disease. *PLOS Comp. Biol.*, 2, 2006.

[47] A. Tovchigrechko and I.A. Vakser. How common is the funnel-like energy landscape in protein-protein interactions. *Prot. Sci.*, 10:1572–1583, 2001.

[48] J. Wang, K. Zhang, H. Lu, and E. Wang. Quantifying the Kinetic Paths of Flexible Biomolecular Recognition. *Biophys. J.*, 2006.

[49] H. J. Dyson and P.E. Wright. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell. Biol.*, 6:197–208, 2005.

[50] F. U. Hartl and M. Hayer-Hartl. Molecular Chaperones in the Cytosol: from Nascent Chain to Folded Protein. *Science*, 295:1852–1858, 2002.

[51] Y.-C. Tang, H.-C. Chang, A. Roeben, D. Wischnewski, N. Wischnewski, M.J. Kerner, F. U. Hartl, and M. Hayer-Hartl. Structural Features of the GroEL-GroES Nano-Cage Required for Rapid Folding of Encapsulated Protein. *Cell*, 125:903–914, 2006.

[52] C.M. Dobson. Protein folding and misfolding. *Nature*, 426:884–890, 2003.

[53] M. Stefani and C.M. Dobson. Protein aggregation and aggregate toxicity: new insights into protein folding, misfolding diseases and biological evolution. *J. Mol. Med.*, 81:678–699, 2003.

[54] F. Chiti, M. Stefani, N. Taddei, G. Ramponi, and C. M. Dobson. Rationalization of mutational effects on protein aggregation rates. *Nature*, 424:805–808, 2003.

[55] C.W. Bertoncini, Jung Y.-S., C.O. Fernandez, W Hoyer, C. Griesinger, T.M. Jovin, and M. Zweckstetter. Release of long-range tertiary interactions potentiates aggregation of natively unstructured α-synuclein. *Proc. Natl. Acad. Sci.*, 102:1430–1435, 2005.

[56] M. Bisaglia, A. Trolio, M. Bellanda, E. Bergantino, L. Bubacco, and S. Stefano Mammi. Structure and topology of the non-amyloid-beta component fragment of human α-synuclein bound to micelles: Implications for the aggregation process. *Prot. Sci.*, 15:1408–1416, 2006.

[57] L. Chiche, L.M. Gregoret, F.E. Cohen, and P.A. Kollman. Protein model structure evaluation using the solvation free energy of folding. *Proc. Natl. Acad. Sci.*, 87:3240–3243, 1990.

[58] D. Eisenberg and A. D. McLachlan. Solvation energy in protein folding and binding. *Nature*, 319:199 – 203, 1986.

[59] J. Novotny, A. A. Rashin, and R. E. Bruccoleri. Criteria that discriminate between native proteins and incorrectly folded models. *Proteins*, 4:19–30, 1988.

[60] B. Lee and F.M. Richards. The interpretation of protein structure: estimation of static accessibility. *J. Mol. Biol.*, 119:537–555, 1971.

[61] R. Sayle and E.J. Milner-White. Rasmol: Biomolecular graphics for all. *Tr. Bioch. Sci.*, 20:374–375, 1995. ftp://ftp.dcs.ed.ac.uk/pub/rasmol/.

[62] W.L. Koltun. Precision space-filling atomic models. *Biopolymers*, 3:665–679, 1965.

[63] M.L. Connolly. Analytical molecular surface calculation. *J. Appl. Cryst.*, 16:548–558, 1983.

[64] F. Eisenhaber and P. Argos. Improved strategy in analytic surface calculation for molecular systems: handling of singularities and computational efficiency. *Journal of Computational Chemistry*, 14:1272–1280, 1993.

[65] M. Totrov and R. Abagyan. The contour buildup algorithm to calculate the analytical molecular surface. *J. Struct. Biol.*, 116:138–143, 1996.

[66] A. Shrake and J.A. Rupley. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.*, 79:351–371, 1973.

[67] Y. N. Vorobjev and J. Hermans. SIMS: computation of a smooth invariant molecular surface. *Biophys. J.*, 73:722–732, 1997.

[68] R.M. Jackson and M. J. E. Sternberg. A Continuum Model for Protein-Protein Interactions: Application to the Docking Problem. *J. Mol. Biol.*, 250:258–275, 1995.

[69] R.J. R. J. Zauhar. The incorporation of hydration forces determined by continuum electrostatics into molecular mechanics simulations. *Journal of Computational Chemistry*, 1991.

[70] F. Fraternali and W. F. van Gunsteren. An Efficient Mean Solvation Force Model for Use in Molecular Dynamics Simulations of Proteins in Aqueous Solution. *J. Mol. Biol.*, 256:939–948, 1996.

[71] H.S. Ashbaugh, E.W. Kaler, and M.E. Paulaitis. Hydration and Conformational Equilibria of Simple Hydrophobic and Amphiphilic Solutes. *Biophys. J.*, 75:755–768, 1998.

[72] P. Ferrara, J. Apostolakis, and A. Caflisch. Evaluation of a fast implicit solvent model for molecular dynamics simulations. *Proteins*, 46:24–33, 2001.

[73] M.-Y. Shen and K.F. Freed. Long Time Dynamics of Met-Enkephalin: Comparison of Explicit and Implicit Solvent Models. *Biophys. J.*, 82:1791–1808, 2002.

[74] H. Snow, C.d. Nguyen, V.S. Pande, and M. Gruebele. Absolute comparison of simulated and experimental protein-folding dynamics. *Nature*, pages 1–4, 2002.

[75] J.A. Wagoner and N.A. Baker. Assessing implicit models for nonpolar mean solvation forces: The importance of dispersion and volume terms. *Proc. Natl. Acad. Sci.*, 103:8331–8336, 2006.

[76] R. Abagyan, M. Totrov, and D. Kuznetsov. ICM - A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *Journal of Computational Chemistry*, 15:488 – 506, 1994.

[77] F. Cutruzzolà, M. Arese, G. Ranghino, G. van Pouderoyen, G. Canters, and M. Brunori. Pseudomonas aeruginosa cytochrome C551: probing the role of the hydrophobic patch in electron transfer. *Journal of Inorganic Biochemistry*, 88:353–361, 2002.

[78] S.W. Porter, Q. Xu, and A.H. West. Ssk1p Response Regulator Binding Surface on Histidine-Containing Phosphotransfer Protein Ypd1p. *Eukaryot. Cell*, 2:27–33, 2003.

[79] Z. Cheng, Y. Liu, C. Wang, R. Parker, and H. Song. Crystal structure of Ski8p, a WD-repeat protein with dual roles in mRNA metabolism and meiotic recombination. *Prot. Sci.*, 13:2673–2684, 2004.

[80] L. Young, R.L. Jernigan, and D.G. Covell. A role for surface hydrophobicity in protein-protein recognition. *Prot. Sci.*, 3:717–729, 1994.

[81] S. Jones and J.M. Thornton. Analysis of Protein-Protein Interaction Sites using Surface patches. *J. Mol. Biol.*, 272:121–132, 1997.

[82] F. Glaser, D.M. Steinberg, I.A. Vakser, and N. Ben-Tal. Residue Frequencies and Pairing Preferences at Protein-Protein Interfaces. *Proteins*, 43:89–102, 2001.

[83] A.J. Bordner and R. Abagyan. Statistical analysis and prediction of protein-protein interfaces. *Proteins*, 60:353–366, 2005.

2

THE DOUBLE CUBIC LATTICE METHOD

F. Eisenhaber, P. Lijnzaad, P. Argos, C. Sander, and M. Scharf.
The double cubic lattice method: Efficient approaches to numerical integration of surface area and volume, and to dot surface contouring of molecular assemblies.
Journal of Computational Chemistry **16**:273–284 (1995)

The Double Cubic Lattice Method: Efficient Approaches to Numerical Integration of Surface Area and Volume and to Dot Surface Contouring of Molecular Assemblies

FRANK EISENHABER*

Biochemisches Institut der Charité der Humboldt-Universität zu Berlin, Hessische Str. 3–4, D-10115 Berlin-Mitte, Germany

PHILIP LIJNZAAD, PATRICK ARGOS, CHRIS SANDER, and MICHAEL SCHARF

European Molecular Biology Laboratory, Meyerhofstr. 1, Postfach 10.2209, D-69012 Heidelberg, Germany

Received 24 March 1994; accepted 17 August 1994

ABSTRACT

The double cubic lattice method (DCLM) is an accurate and rapid approach for computing numerically molecular surface areas (such as the solvent accessible or van der Waals surface) and the volume and compactness of molecular assemblies and for generating dot surfaces. The algorithm has no special memory requirements and can be easily implemented. The computation speed is extremely high, making interactive calculation of surfaces, volumes, and dot surfaces for systems of 1000 and more atoms possible on single-processor workstations. The algorithm can be easily parallelized. The DCLM is an algorithmic variant of the approach proposed by Shrake and Rupley (*J. Mol. Biol.*, **79**, 351–371, 1973). However, the application of two cubic lattices—one for grouping neighboring atomic centers and the other for grouping neighboring surface dots of an atom—results in a drastic reduction of central processing unit (CPU) time consumption by avoiding redundant distance checks. This is most

*Author to whom all correspondence should be addressed at European Molecular Biology Laboratory, Postfach 10.2209, Meyerhofstrasse 1, D-69012 Heidelberg, Germany. Frank Eisenhaber is a visiting scientist at the EMBL, Heidelberg.

noticeable for compact conformations. For instance, the calculation of the solvent accessible surface area of the crystal conformation of bovine pancreatic trypsin inhibitor (entry 4PTI of the Brookhaven Protein Data Bank, 362-point sphere for all 454 nonhydrogen atoms) takes less than 1 second (on a single R3000 processor of an SGI 4D/480, about 5 MFLOP). The DCLM does not depend on the spherical point distribution applied. The quality of unit sphere tesselations is discussed. We propose new ways of subdivision based on the icosahedron and dodecahedron, which achieve constantly low ratios of longest to shortest arcs over the whole frequency range. The DCLM is the method of choice, especially for large molecular complexes and high point densities. Its speed has been compared to the fastest techniques known to the authors, and it was found to be superior, especially when also taking into account the small memory requirement and the flexibility of the algorithm. The program text may be obtained on request.

Introduction

Surface calculations are widely used in molecular mechanical studies. Surface energy functions are applied to approximate the free energy and enthalpy of solvation.[1-10] A dot representation and a triangulation of a molecular surface is a necessary step in computing the electrostatic reaction field potential via the boundary element method.[11-13] Dot surfaces are used in graphical visualizations of molecules. Folding domains may be located as maxima of the compactness along the protein sequence.[14,15] The contribution of amino acid residues to the solvent accessible surface is used as a parameter in characterization of protein folds and in the generation of three-dimensional (3D) profiles for amino acid sequences.[16-18] Surface shape complementarity is one of the criteria used in predicting docking complexes of ligands with macromolecules.[19]

The surface of a molecule may be defined in different ways. The solvent accessible surface[20] is that part of the surface of a sphere centered at an atom with radius $r_{vdW} + r_{sol}$, where the center of a spherical solvent molecule or probe (r_{sol}) can be placed in contact with the atomic van der Waals sphere (r_{vdW}) without penetrating other atoms. The molecular surface[21] is the envelope of the molecular volume from which solvent is excluded. The molecular surface can often be replaced by the van der Waals surface (accessible surface with zero probe radius), which is easier to calculate.[22]

Accurate methods for calculating molecular surfaces can be grouped into analytical[23-30] and numerical integration approaches. The numerical methods can be characterized by their way of discrete surface approximation: slices of cylindrical surfaces,[20,31] cube compositions,[32-35] and point distributions on atomic spheres[36-41] (the most simple variant). Nearly all algorithms are too slow to be used in a context of multiple invocations, such as in molecular mechanics studies in which the solvation energy of various molecular conformational states is determined.[10]

Recently, two efficient modifications[40,41] of the Shrake and Rupley method[36] have been published. The fastest algorithm, the method of LeGrand and Merz,[40] computes an approximate value of the Shrake and Rupley surface area using a precalculated library of bit strings that code for the burial of atomic surface points by neighboring atoms at some space grid positions. Unfortunately, even at relatively low point density, the library occupies a significant part of the computer memory; for example, 1.4 MBytes for a 256-dot sphere. The method also lacks flexibility, especially if the user alters the distribution or number of surface dots or if a new set of radii for atoms and/or solvent is applied (with larger maximal radii). In such cases, it is necessary to recalculate the library.

The method of Abagyan and Totrov[41] depends on a special dot arrangement, in which the z-coordinate is regularly incremented while the angular position in the z-slice is defined by Fibonacci numbers (Pietr Zielenkiewicz, personal communication). Only dots within a z-axis segment, defined by the projection of the circle of intersection with a neighboring atom onto the z-axis, are checked for burial by the neighbor. However, this point distribution is not very regular, especially at the poles. Abagyan and Totrov rely on a neighbor list based on chemical groups to reduce the number of pairwise atom distance checks.

The double cubic lattice method described in this article overcomes several of the restrictions in the previously described numerical approaches. It is slightly faster than the method of LeGrand and Merz,[40] but the calculated area values are as exact as in the original Shrake and Rupley[36] approach. Additionally, our memory requirement is negligibles in contrast to that of their method. Our technique also does not depend on special properties of the surface dot distribution, as does the method of Abagyan and Totrov.[41] It is simple to implement and to incorporate into existing programs.

The algorithm basically uses an overlayed cubic lattice for grouping spatially close atoms of the molecular assembly. Likewise, another cubic lattice is applied to sort spatially close points of the tessellated atomic sphere. As a result, groups of atoms and points may be immediately recognized as nonoverlapping. Additionally, computer time is saved by the use of the dot product instead of the Euclidean distance for checking overlap.

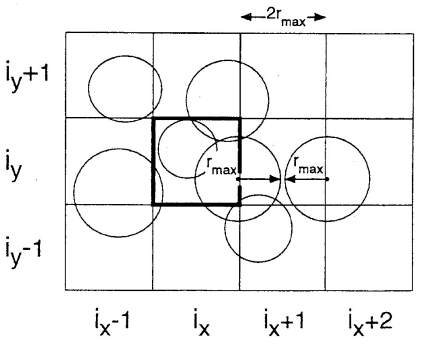

FIGURE 1. The use of a cubic grid for grouping atoms, shown here for the two-dimensional case. Atoms in the central grid cell (i_x, i_y) can only overlap with atoms in the neighboring grid cells $(i_x, i_y \pm 1)$, $(i_x \pm 1, i_y)$, and $(i_x \pm 1, i_y \pm 1)$. Overlap between the large atoms in (i_x, i_y) and $(i_x, i_y \pm 2)$ cannot occur if the grid spacing is chosen to be $2r_{max}$.

Calculation Methods

THE DCLM

The method of artificial grids is a standard approach in computational geometry to group spatially close objects.[41] The cubic lattice is the simplest, allowing the division of the three-dimensional space into disjunct and equal elementary volumes with equal extensions along the three global Cartesian coordinate axes. In the subsequent description of mathematical details, the following notations are used: $\mathbf{a}_i = (x_i, y_i, z_i)$ for the coordinate vector of atoms $i = 1, \ldots, N$ in a Cartesian coordinate space; r_i for the sum of the van der Waals and solvent radii for each respective atom (subsequently referred to as the atomic radius); r_{max}, the maximum of these radii, d_{ij} for the distance between atoms i and j; and $\mathbf{p}_k$ for the coordinate vector of dots $k = 1, \ldots, m$ on a unit sphere centered at the origin of the global coordinate system. The tessellation of the unit sphere is given by an m – tuple $T = (\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_m)$.

The list of neighboring atoms is obtained with the following procedure. The molecular assembly under consideration is placed into a single rectangular box with edge lengths being an integer multiple of $2r_{max}$. This box is subdivided into grid cells by a cubic grid with the spacing $2r_{max}$. As a result, each atom has neighbors only in its own grid cell and in its (maximally 26) neighboring grid cells (Fig. 1). The computational steps involved are as follows:

1. Determination of the minimum and the maximum of the x-, y-, and z-coordinates of all atoms
2. Calculation of the origin of the large rectangular box (x_o, y_o, z_o) and the number of grid cells along each coordinate axis (b_x, b_y, b_z)
3. Assignment of the corresponding subbox number to each atom (i.e., if some atom lies in box (i_x, i_y, i_z) with $0 \le i_x < b_x$, $0 \le i_y < b_y$, and $0 \le i_z < b_z$, then its box number is equal to $i_x + (i_y \cdot b_x) + (i_z \cdot b_x \cdot b_y)$
4. Sorting of the atoms in accordance with box numbers.

The number of boxes occupied by atom centers is small compared with the total number of atoms. Therefore, special techniques of sorting with a small number of integer keys (with order of N operations) can be employed. Step 4 is elegantly executed with an algorithm similar to that published by Sedgewick[42] or, equivalently, using linked lists of atoms rooted in the grid cells to which the atoms belong.[43] The complete neighbor list is computed within a few hundredths of a second even for a protein with more than 3000 atoms.

The idea of neighbor lists based spatial grouping as an alternative to the chemical group method has already been suggested by other authors (see refs. 43 and 44 and references therein). It is generally concluded that the grid cell approach for calculating nonvalent interaction lists is efficient only for systems with about 600 atoms or more. Our neighbor list calculation algorithm is also suitable for much smaller systems. Obviously, the improved performance of our algorithm compared with that of Yip and Elber[44] is explained by two computational details. We use (1) variable box numbers with a fixed grid spacing equal to $2r_{max}$, the cutoff for surface interactions (a relatively small value, about 7 Å); and (2) an effective sorting approach. Cubic lattices have also been applied to approximate electrostatic forces in macromolecules.[45] For long-range interactions, the grid cell approach is less effective, and large memory requirements may cause problems.[43]

Later we will discuss aspects of unit sphere tesselations. The special properties of a point distribution $T = (\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_m)$ on a unit sphere do not affect the present algorithm; however, the distribution should be even as possible to minimize integration error. The points $\mathbf{p}_i$ of the unit sphere sampling the surface of the atom under consideration must be checked for occlusion by neighboring atoms. To speed up this step, we employ a second lattice contained in a cube that is inscribed by the unit sphere (Fig. 2). This box is subdivided into M^3 equal small cubes. A list of surface dots per box is calculated in a way analogous to the previously given recipe (steps 1 through 4).

The information obtained from the second lattice may be used in various ways for surface calculation. The first variant, instigated by the method of Abagyan and Totrov,[41] directly cycles over the elementary cubes containing the dots. In the second approach, the unit sphere lattice is only implicitly used. A third modification investigated by us will be outlined.

Variant I

To computate the surface of atom i, a first loop is performed over all atoms $j \neq i$ from its own grid cell and all atoms j in the neighboring boxes. For a given atom j, the intervals $[X_l', X_u']$, $[Y_l', Y_u']$, and $[Z_l', Z_u']$, which are the projections of the sphere j onto the axes of the local grid (Figs. 2 and 3) are calculated. These intervals define a rectangular

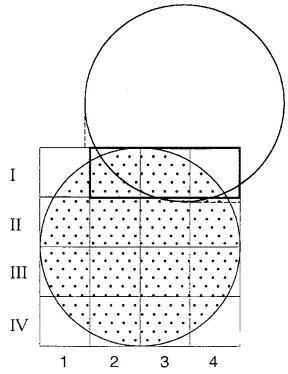

FIGURE 2. The use of a cubic grid for grouping surface dots, shown here for the two-dimensional case. In variant I, the fact is used that only dots lying in the area of overlap can get buried by the neighboring atom. Here, only dots lying in the boxes I2–4 have to be checked for occlusion (bold box). In a simpler scheme, the full extent of the neighboring atom is taken instead of the extent of the intersection circle. Here, that would result in the need to check the dots in the boxes I1 and II1–4 (dashed box). In variant II, the grid is employed to sort all dots according to their spatial vicinity (i.e., all dots in one and the same elementary cube are put successively into the array of dot coordinates).

box that contains the only points on atom i that can be occluded by atom j. For the x-axis, we find

$$X_l' = \max\left(\frac{x_j - x_i - (r_i - r_j)}{2r_i}, 0\right) \quad \text{and} \quad \text{(1a)}$$

$$X_u' = \min\left(\frac{x_j - x_i + (r_i + r_j)}{2r_i}, 1 - \frac{1}{M}\right) \quad \text{(1b)}$$

The expressions for the other directions are similar. The function

$$g(X) = \text{int}(M \cdot X) \quad \text{(1c)}$$

translates the interval boundaries into grid position where int represents the nearest smaller integer. For point densities higher than about 400 points per sphere (M larger than 4), it is advantageous to take only the projection of the intersection circle onto the three coordinate axes as limits for the respective intervals $[X_l, X_u]$, $[Y_l, Y_u]$, and $[Z_l, Z_u]$ in the same way, as Abagyan and Totrov

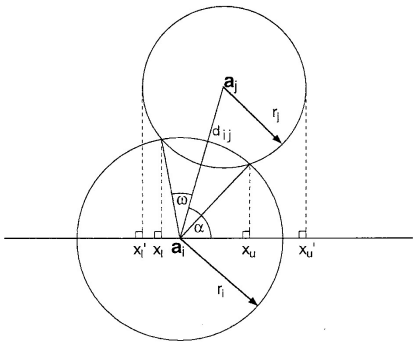

FIGURE 3. Construction of the projection of an atom's extent, or the extent of the intersection circle, onto an axis of the grid. Shown here for the x-axis. $\mathbf{a}_{i,j}$, $r_{i,j}$, and d_{ij} as usual; $X'_{l,u}$ the lower and upper limit of the projection of atom j's extent onto the x-axis; $X_{l,u}$ the lower and upper limit of the projection of the intersection circle onto the x-axis; a the angle of the vector $\mathbf{a}_j - \mathbf{a}_i$ with the positive x-axis; u the radius of the intersection circle, expressed as an arc over the surface of the unit sphere. In the text, $X'_{l,u}$ and $X_{l,u}$ are expressed as fractions of r_i.

did for one axis[40] (Figs. 2 and 3). The interval boundaries $[X_l, X_u]$ then become

$$X_l = \begin{cases} 0 & \text{if } \cos(\alpha) \le -\cos(\omega) \\ \dfrac{\cos(\alpha + \omega) + 1}{2} & \text{otherwise} \end{cases} \tag{1a'}$$

and

$$X_u = \begin{cases} 1 - \dfrac{1}{M} & \text{if } \cos(\alpha) \ge \cos(\omega) \\ \dfrac{\cos(\alpha - \omega) + 1}{2} & \text{otherwise} \end{cases} \tag{1b'}$$

Although calculation of one square root per coordinate axis cannot be avoided, the computational costs are mitigated by an additional reduction of surface dot checks. The cosines of $\alpha + \omega$ and $\alpha - \omega$ are calculated via

$$\cos(\alpha)\cos(\omega) = \frac{x_j - x_i}{r_i} C \quad \text{and} \tag{2a}$$

$$\sin(\alpha)\sin(\omega) = \frac{1}{r_i}\sqrt{\left(\frac{r_i^2}{d_{ij}^2} - C^2\right)\left((y_j - y_i)^2 + (z_j - z_i)^2\right)} \tag{2b}$$

where

$$C = \frac{d_{ij}^2 + r_i^2 - r_j^2}{d_{ij}^2} \tag{2c}$$

For each nonempty grid cell, we loop over all dots that are not yet buried. The occlusion of the surface dot $\mathbf{p}_k$ by the neighboring atom j is not tested by the squared distance as usually done, but via the dot product, which is slightly faster to compute. Dot $\mathbf{p}_k$ is occluded by atom j if

$$(\mathbf{a}_j - \mathbf{a}_i) \cdot \mathbf{p}_k > \frac{d_{ij}^2 + r_i^2 - r_j^2}{2r_i} \tag{3}$$

where the value on the right side has to be calculated only once for each neighboring atom j. It was found empirically that M is a good choice if

$$M^3 \le m/2 < (M+1)^3 \tag{4}$$

where m is the number of points per sphere. In this case, the computer time consumption is low for a wide range of m, but it may not be the optimal selection of M for a special value of m.

Variant II

Here the loop over the dots k on the sphere of atom i precedes the loop over the neighboring atoms j. The rationale is that most of the occluded surface dots are buried by a few close atoms. If dot $\mathbf{p}_k$ is occluded by neighboring atom j, the number j is stored and it is checked whether the next dot $\mathbf{p}_{k+1}$ is buried by the same atom j. In this variant, it is necessary to precalculate a temporary list of neighbors j overlapping atom i on the basis of the neighbor list obtained with the first cubic grid. Also, arrays with the difference vectors $\mathbf{a}_j - \mathbf{a}_i$ and the corresponding dot products [see eq. (3)] must be prepared.

This algorithm is sensitive to the order in the m-tuple $T = (\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_m)$ of point vectors. Given that atom j is occluding dot $\mathbf{p}_k$, it would be desirable to have all other dots covered by atom j following $\mathbf{p}_k$ in the m-tuple T. This can be partly fulfilled by reordering the point vectors in such a manner that $\mathbf{p}_k$ is followed by a spatially close dot $\mathbf{p}_{k+1}$ in T. If the algorithm generating the tessela-

tion T does not place the dots in such an order automatically (as in DSSP[46]), the cubic lattice on the unit sphere can be applied and all dots in one elementary cube are placed successively into the corresponding arrays. Additionally, dots from neighboring boxes would similarly follow each other.

If now a dot $\mathbf{p}_k$ is occluded by one neighboring atom j, the likelihood that the same atom will overlap the following dot $\mathbf{p}_{k+1}$ is high. Therefore, it is worthwhile to store the atom j and to check it with the next dot. If atom j does not occlude dot $\mathbf{p}_{k+1}$, only then is the iteration over the temporary neighbor list restarted. As a result, the central loop of the corresponding program is remarkably short. Our statistics for the solvent accessible surface of compact protein structures show that about 80% of all dots are recognized as buried after checking ony one neighboring atom j. The worst case is an accessible dot for which a complete loop over all neighbors j has to be made.

In this variant, the grid spacing of the unit sphere depends mainly on the depth of overlap of a good neighbor and not on the dot density. Good performance can be achieved for $M = 3$, 4, 5, and 6; $M = 4$ is a generally acceptable value for all dot densities and protein structures studied.

Variant III

It is attractive to loop, for a given atom i, over all elementary cubes occupied with dots in the unit sphere grid and then cycle over all neighboring atoms j. The corner of the elementary cube with the largest distance from the center of atom i can be rapidly determined by considering each coordinate axis individually. For the x-axis, the coordinate difference d_x between the center of atom j and the farthest corner is

$$d_x = \max\left(\frac{x_j - x_i}{r_i} - X_l, \frac{x_j - x_i}{r_i} - X_l - \Delta\right) \quad (5)$$

where X_l and $X_l + \Delta$ are the two possible x-coordinates of the eight vertices of cube l and Δ is the grid spacing. If this corner is occluded by atom j, all dots in cube l are also not accessible. Similarly, an elementary cube is not occluded by any atom j if its circumscribed sphere has no overlap with any neighbor, and consequently all dots in the corresponding cube are accessible.

The status of most of the dots is determined by one of the two conditions just described.

Relative Performance of the Variants

We also investigated several other combinations of the previously described algorithmic elements. All variants proved faster than previously published algorithms. If implemented on a UNIX workstation (SGI or DEC), variant II is faster than any other of our variants. Algorithm II outperformed variant I by a factor of almost 2 and variant III by a factor of about 1.5. All results presented in this article have been obtained with variant II.

We think that the preferences may be different for other machine architectures and/or compilers. The burden of conditional jumps, the ability to execute integer and float operations in parallel, and the size of fast cache memory will be decisive for this choice.

NOTES ON PARALLELIZATION

The time-consuming part of the calculation is the check of the dots on a given atomic sphere for burial by neighboring atoms. The calculations for atoms i are independent from each other and can be placed on independent processors. The use of the *m_fork* utilities (Sequent Computer Systems parallel programming primitives) as available on SGI multiprocessor machines is a cheap possibility. The reduction in computation time is expected to be almost linear with the number of processors.

COMPUTATION OF SURFACE, VOLUME, AND COMPACTNESS

The surface area A is obtained by

$$A = 4\pi \sum_i r_i^2 \frac{m_{\mathrm{acc}}(i)}{m} \quad (6)$$

where $m_{\mathrm{acc}}(i)$ is the number of dots on atom i not occluded by neighboring atoms and the summation is over all atoms in the molecule. The accuracy may be improved if the surface dots are weighted by area in accordance with the Voronoi subdivision of the unit sphere for a given tesselation.

The molecular volume V is calculated through the Gauss-Ostrogradskii theorem,

$$\int_V \nabla \cdot \mathbf{F}(\mathbf{r})\, dV = \oint_S \mathbf{F}(\mathbf{r}) \cdot dS$$

with $\mathbf{F}(\mathbf{r}) = \mathbf{r}$, the coordinate vector. As a result, the volume is equal to

$$V = \frac{4\pi}{3m} \sum_i \left[r_i^2 \mathbf{a}_i \cdot \left(\sum_{k(i)} \mathbf{p}_k \right) + r_i^3 m_{\text{acc}}(i) \right] \quad (7)$$

where all dots are assigned equal weights. The second summation is carried out only over all accessible dots $\mathbf{p}_k$ on atom i. The numerical error in the volume calculation is also affected by the average length of the vectors $\mathbf{a}_i$. This influence can be suppressed by placing the global origin at the geometrical center of the molecular assembly under study. The volume in eq. (7) is the CPK (Corey-Pauling-Koltun) or van der Waals volume if the atom radii are not incremeted by the probe radius. This volume is a good approximation to the solvent-excluded volume defined by Connolly[47] for low molecular compounds. With atom radii including the probe radius, eq. (7) yields the volume inside the solvent accessible surface used in compactness calculations.

The compactness z of a molecule is defined as the solvent accessible surface area divided by the minimum possible surface area, which is that of a sphere of equal volume.[15] Thus,

$$Z = A \cdot (36\pi \cdot V^2)^{-1/3} \quad (8)$$

TESSELATION OF THE UNIT SPHERE

The spherical point distribution to be used for a Shrake and Rupley type of surface calculation has to be as even as possible. Subdivisions based on inscribed regular polyhedra are natural candidates. Tesselations have been frequently studied, with applications ranging from architecture (geodesic domes) to biology (structure of viral coats).[48]

An icosahedral tesselation produces $10\tau + 2$ vertices (20τ faces, 30τ edges), where τ is of the form

$$\tau = b^2 + bc + c^2$$

Three classes of tesselations can be distinguished: $c = 0$ (class I); $b = c$ (class II); and $b \neq c$ (class III, skewed tesselations).

The quality of a tesselation, here defined as the ratio of the maximum and minimum edge lengths of the Delaunay triangulations it generates, depends on the sophistication of the algorithm. A yardstick for the quality is the case of the tesselation frequency tending to infinity. The five triangles surrounding the icosahedral poles then form regular pentagons, resulting in an edge length ratio of $2\sin(36^\circ) \approx 1.1756$.

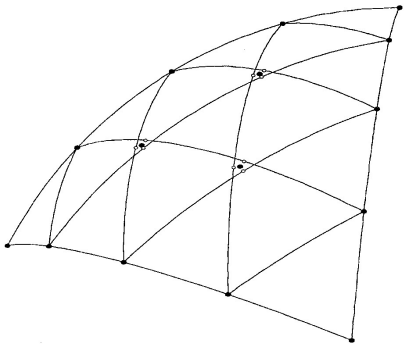

FIGURE 4. Tessellation of an icosahedral face with frequency 4. Equiangular subdivision of the icosahedral edges does not yield coincident equivalent grid points inside the face. The average point (filled circles) of three equivalent grid points (open circles) is taken instead.

The simplest way to subdivide a spherical triangle is through projection onto the unit sphere of a triangular grid lying on a triangular face (gnomonic projection.[49] This, however, results in poor quality, with a ratio of longest to shortest arc of about 1.5 because the subdivision is equal with respect to chords but not spherical arcs.

We have cast this way of subdivision in spherical terms, by covering the spherical triangle with segments of great circles that connect corresponding subdivisions of the edges of the spherical triangle (Fig. 4). These segments form a grid similar to the triangular grid previously mentioned. However, these segments, and their subsequent subdivisions into the spherical triangular grid points, do not coincide generally (Fig. 4). This problem was solved by always taking the average of the three resulting equivalent grid points. When such a method of tesselating a face is applied to the icosahedron or pentakisdodecahedron (32 vertices, 90 edges, 60 faces), one obtains class I or II icosahedral tesselations, respectively.

We found the tesselation quality of this procedure for any frequency always to be better than the limiting ratio of 1.1756. This is better than any of the methods currently in use.[39,46,50,51] Another

advantage is the good control over the number of points produced, as one can choose $10b^2 + 2$ points (class I tesselations) or $30b^2 + 2$ points (class II tesselations), with b being the tesselation frequency (which can be any positive integer).

The procedure used to triangulate the point distributions is based on a so-called greedy planar triangulation[52] but extended to deal with points on the unit sphere. It generates a Delaunay triangulation from which edge lengths and areas of faces and Voronoi regions are calculated. Areas are those of spherical polygons.

TABLE I.
Comparison of Numerical and Analytical Surface Area and Volume Calculations and the Influence of Point Density.

	Analyt.	122	362	642	1002	1472
		Ia. 4PTI (454 heavy atoms, 58 residues)				
t	2.97	0.50	0.91	1.36	1.86	2.54
A	3973.8	3961.4	3971.8	3967.9	3974.1	3975.7
ΔA	0.0	13.4	2.0	5.9	0.3	1.9
ΔA_{atom}	0.0	4.1	1.5	1.0	0.6	0.6
V	11915.3	11871.3	11923.7	11885.7	11911.1	11911.9
ΔV	0.0	44.0	8.4	29.6	4.2	4.6
Z	1.575	1.574	1.574	1.575	1.576	1.576
		Ib. 3FXN (1073 heavy atoms, 138 residues)				
	Analyt.	122	362	642	1002	1472
t	8.15	1.29	2.21	3.15	4.31	5.63
A	6943.8	6968.3	6933.4	6944.4	6939.1	6943.3
ΔA	0.0	24.5	10.4	0.6	4.7	0.5
ΔA_{atom}	0.0	4.5	1.5	1.3	0.7	0.6
V	26350.0	26373.6	26281.9	26370.6	26338.8	26352.3
ΔV	0.0	23.6	69.1	20.6	11.2	2.3
Z	1.621	1.626	1.622	1.621	1.621	1.621
		Ic. 1TIM (3740 heavy atoms, 492 residues, 2 chains)				
	Analyt.	122	362	642	1002	1472
t	32.1	5.23	8.18	11.55	15.25	20.51
A	20002.8	19970.9	19997.1	19998.7	20012.2	19997.0
ΔA	0.0	31.9	5.7	4.1	9.4	5.8
ΔA_{atom}	0.0	4.2	2.1	1.3	1.0	0.7
V	89100.7	89153.7	88972.7	89029.1	89121.0	89067.3
ΔV	0.0	53.0	128.0	71.6	21.0	33.4
Z	2.073	2.069	2.075	2.074	2.074	2.073

Results of the program NSC for selected protein structures (Tables Ia–c) are presented. Coordinates of all ATOMS records were taken from the corresponding entries of the Brookhaven Protein Data Bank.[53,54] The atom radii were taken from Eisenberg and McLachlan.[3] The solvent radius was set equal to 1.4 Å.

The time t (in seconds) is the CPU time for calculating the surface (on a single R3000 processor of a SGI / 4D 480 as in ref. 30, about 5 MFLOP). The time values here are directly comparable with $t_{ana,3}$ in Tables 1 and 2 in ref. 30. Both the numerical method and the analytical routine (ASC[30]) show a significantly improved performance. For the numerical surface computation, we applied variant II (see the Calculation Methods section).

A (in Å^2) is the solvent accessible surface area for the whole structure. ΔA denotes the absolute value of the difference between the value of A and the analytically calculated surface area using the program ASC. ΔA_{atom} is the maximum of the absolute deviation of the surface of a single atom and its analytically computed area. The volume V (in Å^3) inside the solvent accessible surface, the absolute deviation from the analytically calculated volume, and the compactness Z [eq. (4)] are also presented.

To assess numerical error, the analytical results for surface area (from ASC[30]) and volume are listed in the second column. The analytical volume was computed with the program PQMS[46] using atom radii incremented by the solvent radius and a zero probe radius. In this case, cusps do not occur and the volume calculation is completely analytical.

The following columns present the numerical results for several dot densities of the atomic spheres; tessellations 122 and 1472 are based on the dodecahedron (class II), whereas the remaining are icosahedral tessellations (class I): (a) 4PTI (454 heavy atoms, 58 residues); (b) 3FXN (1073 heavy atoms, 138 residues); (c) 1TIM (3740 heavy atoms, 492 residues, 2 chains).

Results

A computer program NSC (numerical surface calculation) was written in standard C and tested on several types of UNIX workstations (SGI, DEC, SUN). For file input and command interpretation, it was attached to the shell of ASC.[30] Best optimization results were obtained with the compiler *gcc* distributed by the Free Software Foundation (version 2.5.7, options *-O2 -ffast-math -finline-functions -funroll-all-loops*). NSC selects among the available unit sphere tesselations the one with the smallest number of dots above the density required by the user. The calculation results for selected protein structures are presented in Tables Ia–c and II.

The accuracy of our numerical surface integration technique was checked against the analytical computation of the surface area for the entire molecule and also on a per-atom basis. For this purpose, the program ASC[30] was used. The accuracy is improved with increased dot density (Tables Ia–c). More than 600 dots per unit sphere ensure an accuracy better than 1.5 Å^2 for the surface area for every atom which is about 1% of the solvent accessible surface area of an isolated atom.

The quality of the unit sphere tesselation is decisive for the spatial invariance[38] (independence of the surface area on the orientation of the molecule with respect to the reference frame of the point distribution). Our polyhedral point distributions yield spatially invariant surface area values because the accuracy of surface area per atom depends only on point density for various protein structures (differently oriented in their crystals; see Tables Ia–c).

Interestingly, especially for higher point densities, the surface area results are practically the same, whether the polyhedral or the Zielenkiewicz point distribution[41] is applied. This is surprising because the tesselation quality of the latter is only between 2.5 and 3. The edge length distribution of the corresponding Delaunay triangulation is shown in Figure 5a and can be seen to be rather poor. Nonetheless, the face areas and the Voronoi areas (for our purposes more relevant) are well distributed (see Figs. 5b and 5c). Consequently, the Zielenkiewicz distribution can safely be applied for the calculation of surface areas of entire molecules, but for most other purposes (e.g., for generating a dot surface to be used in trianglation), a polyhedral distribution is preferred. It can also be seen from Figure 5c that the introduction of point weights for the polyhedral tesselation would not significantly improve the area accuracy if the dot density is low (less than 1000).

The computation speed is influenced by the number of points per atomic sphere, by the number of atoms, and by the conformation of the molecule. In more compact molecules, each atom has a larger number of neighboring atoms (Table II). Our numerical routine is normally faster than ASC[30] for point densities below 1000. Because of the rapid neighbor list calculation and a reduced number of distance checks between atoms and surface dots, our method is faster than the approach of Abagyan and Totrov; for example, NSC needs only 34% of their CPU time for flavodoxin (atom coordinates of entry 3FXN in the Protein Data Bank). The calculation of the solvent accessible surface area of the crystal conformation of bovine pancreatic trypsin inhibitor (entry 4PTI of the Brookhaven Protein Data Bank, 362 point sphere for all 454 nonhydrogen atoms) takes less than 1 second (Table Ib). The same CPU time was achieved by the algorithm of Merz and LeGrand[40] on a comparable machine, but with a lower point density. It should also be noted that NSC was used inside the command shell of ASC and, as a result, the access to coordinates and radii was not optimized by the compiler together with the numerical surface calculation procedure. This is a realistic situation because surface calculations are often only one of the computational steps in a complex

TABLE II.
Influence of Conformation.

	1CRN		$CRN_{extended}$	
	Analyt.	642	Analyt.	642
t	1.96	0.96	0.91	0.80
A	2988.3	2995.8	5813.5	5815.8
ΔA	0.0	7.5	0.0	2.3
ΔA_{atom}	0.0	1.1	0.0	1.1
V	8687.3	8703.5	10829.8	10819.2
ΔV	0.0	16.2	0.0	10.6
Z	1.462	1.464	2.456	2.458

The data show the dependency of computational efficiency on structural compactness. Crambin is a small protein with 327 heavy atoms and 46 residues. The data are presented for the PDB version (1CRN[53,54]) and for a fully extended conformation ($CRN_{extended}$). The symbols are assigned as in Table I. The atom coordinates for the extended conformation of crambin were obtained using the program package ICM.[41]

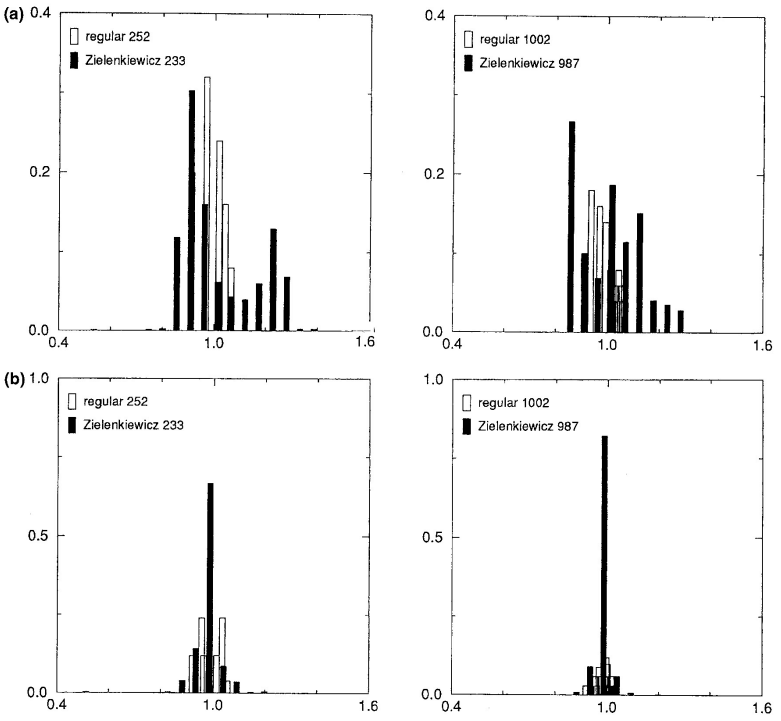

FIGURE 5. Comparison of the quality of point distributions at different numbers of points per sphere. The facial and Voronoi areas and edge lengths refer to those in a Delaunay triangulation[52] of the distributions. The plots represent histograms of fractional deviations from the ideal value. The data were grouped in 20 bins of equal width; plotted vertically is the occurrence as fraction of the total. (a) Edge lengths. The ideal edge length L' is calculated under the assumption that the unit sphere is uniformly covered with equilateral spherical triangles of arc length L. For the area ε of one such triangle, we can write

$$\varepsilon = \frac{4\pi}{F} = 4 \arctan\left(\sqrt{\tan\left(\tfrac{3}{4}L\right)\tan^3\left(\tfrac{1}{4}L\right)}\right)$$

(formula of L'Huilier from spherical trigonometry), where F is the number of triangular faces. After solving for L, the idea edge length is given by $L' = 2\sin(L/2)$. (b) Face areas. The ideal value is $4\pi/F$, where F is the number of faces per unit sphere. (c) Areas of Voronoi regions. The ideal value is $4\pi/m$, where m is the number of dots per unit sphere.

context. With a one-file program, even better performance can be expected. For a molecule of about 1000 atoms and a medium dot density per atomic sphere, the calculations are in fact done interactively even on a normal workstation.

The volume values V calculated with our routine differ less than 3% from the analytically calculated volumes (program PQMS),[46] beginning with a point density of 600 dots per atomic sphere (Tables Ia–c). Similarly, the compactness values computed with our routine are identical to those calculated by Zehfus[15] with an accuracy of 2%.

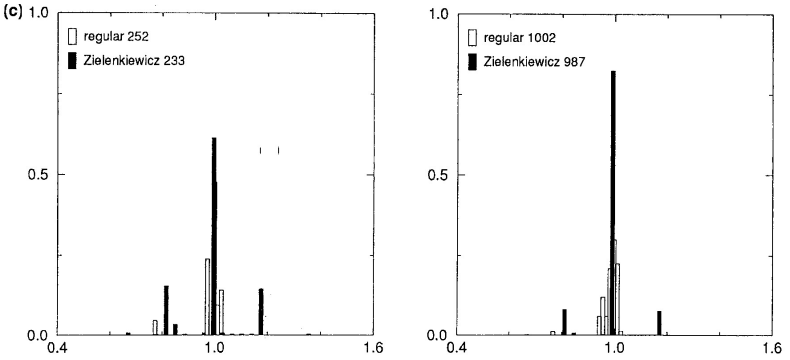

FIGURE 5. (Continued)

Discussion

All surface calculation procedures have advantages and deficiencies. The choice of the best method for a certain purpose depends on many aspects; for example, the desired accuracy of the surface area (overall value or area per atom), computational speed, memory requirements, algorithmic complexity, available time for programming efforts, and the context of the calculation.

An analytical algorithm[30] is to be preferred if the accuracy of the surface area per atom is important, such as in optimization protocols. The noise of numerical surface evaluation may render the total energy function misbehaved to numerical optimization routines. Analytic surface computation is also advantageous if it is desirable to obtain at low additional computational costs accurate analytical derivatives of the surface area (with respect to atom coordinates) or the connectivity of surface pieces.[47,55] The computation speed is not a reason to reject the analytical algorithm. The program ASC[30] (version 2.0, February 1994) calculates the surface area more than two times faster than previously described (especially for large and compact molecules) as a result of several algorithmic innovations (compare t in the columns for analytical results of Tables Ia–c and II with $t_{ana,3}$ in Tables 1 and 2 of ref. 30). The need for extensive and sophisticated programming, however, is a drawback of the analytical method.

A numerical algorithm as presented here is advantageous (1) if the accuracy of surface area values in the range of ± 0.5 to 3.0 Å^2 per atom (depending on point density) is sufficient and (2) if it is desired to determine at almost no additional computational cost the values of the volume (via the Gauss-Ostrogradskii theorem; see the Calculation Methods section) and of the compactness or the coordinates of dots representing the surface of the molecule. Our polyhedral dot distributions are well suited for generating dot surfaces that may be further used for graphics, triangulation, boundary element calculations (dot selection via the spoke method[13]), and geometrical docking. The algorithms described in this article will be applied in new releases of the programs DSSP[46] and Melc.[13]

The code of the subroutine NSC in standard C is available for distribution; the request should be directed to the authors by regular post or electronic mail: Eisenhaber@EMBL-Heidelberg.DE on Internet. The routine NSC is also incorporated into the program ASC[30] (beginning with version 2.0) and can be invoked with the commands "nsc", "nsc_vol", and "nsc_dots". The package ASC is available via FTP (send a request for details).

Acknowledgments

The authors are grateful to C. J. Kitrick for suggestions regarding tessellation of the unit sphere and for providing his sphere tessellation programs to us. F. E. is grateful to R. Abagyan for generously giving insight into details of his program package ICM[40] and thanks A. Juffer for stimulating discussions and C. Frömmel for encourage-

ment. The authors acknowledge financial support from the German Bundesministerium für Forschung und Technologie (grant FG5-1075 to P.A.) and from the Fund Wissenschaftler-Integrationsprogramm (grant 020386/B to F.E.) jointly administered by the East German Länder and the government of the Federal Republic of Germany.

References

1. W. Hasel, T. Hendrickson, and W. C. Still, *Tetr. Comp. Method.*, **1**, 103 (1988).
2. W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson, *J. Amer. Chem. Soc.*, **112**, 6127 (1990).
3. D. Eisenberg and A. D. McLachlan, *Nature*, **319**, 199 (1986).
4. L. Wesson and D. Eisenberg, *Protein Science*, **1**, 227 (1992).
5. C. A. Schiffer, J. W. Caldwell, R. M. Stroud, and P. A. Kollman, *Protein Science*, **1**, 396 (1992).
6. C. A. Schiffer, J. W. Caldwell, P. A. Kollman, and R. M. Stroud, *Molecular Simulation*, **10**, 121 (1993).
7. B. v. Freyberg and W. Braun, *J. Comp. Chem.*, **14**, 121 (1993).
8. B. v. Freyberg, T. J. Richmond, and W. Braun, *J. Mol. Biol.*, **233**, 275 (1993).
9. T. Ooi, M. Ootobake, G. Nemethy, and H. A. Scheraga, *Proc. Natl. Acad. Sci.*, **84**, 3086 (1993).
10. K. C. Smith and B. Honig, *Proteins*, **18**, 119 (1994).
11. R. J. Zauhar and R. S. Morgan, *J. Comp. Chem.*, **11**, 603 (1990).
12. B. J. Yoon and A. M. Lenhoff, *J. Comp. Chem.*, **11**, 1080 (1990).
13. A. H. Juffer, E. F. F. Botta, B. A. M. van Keulen, A. van der Ploeg, and H. J. C. Berendsen, *J. Comp. Phys.*, **97**, 144 (1991).
14. M. Zehfus and G. D. Rose, *Biochemistry*, **25**, 5759 (1986).
15. M Zehfus, *Proteins*, **16**, 293 (1993).
16. J. Bowie, R. Luthy, and D. Eisenberg, *Science*, **253**, 164 (1991).
17. R. Luthy, J. Bowie, and D. Eisenberg, *Nature*, **356**, 83 (1992).
18. R. Abagyan, D. Frishman, and P. Argos, *Proteins*, **19**, 132 (1994).
19. J. Cherfils and J. Janin, *Curr. Op. Struct. Biol.*, **3**, 265 (1993).
20. B. Lee and F. M. Richards, *J. Mol. Biol.*, **55**, 379 (1971).
21. F. M. Richards, *Ann. Rev. Biophys. Bioeng.*, **6**, 151 (1977).
22. D. C. Rees and G. M. Wolfe, *Protein Science*, **2**, 1882 (1993).
23. M. L. Connolly, *J. Appl. Cryst.*, **16**, 548 (1983).
24. M. L. Connolly, *J. Appl. Cryst.*, **18**, 499 (1985).
25. T. J. Richmond, *J. Mol. Biol.*, **178**, 63 (1984).
26. K. D. Gibson and H. A. Scheraga, *Mol. Physics*, **62**, 1247 (1987).
27. K. D. Gibson and H. A. Scheraga, *Mol. Physics*, **64**, 641 (1988).
28. L. R. Dodd and D. N. Theodorou, *Mol. Physics*, **72**, 1313 (1991).
29. G. Perrot, B. Cheng, K. D. Gibson, J. Vila, K. A. Palmer, A. Nayeem, B. Maigret, and H. A. Scheraga, *J. Comp. Chem.*, **13**, 1 (1992).
30. F. Eisenhaber and P. Argos, *J. Comp. Chem.*, **14**, 1272 (1993).
31. T. J. Richmond and F. M. Richards, *J. Mol. Biol.*, **119**, 537 (1978).
32. J. J. Müller, *J. Appl. Cryst.*, **16**, (1983).
33. M. Y. Pavlov and B. A. Fedorov, *Biopolymers*, **22**, 1507 (1983).
34. A. Y. Meyer, *J. Comp. Chem.*, **9**, 18 (1988).
35. H. R. Karfunkel and V. Eyraud, *J. Comp. Chem.*, **10**, 628 (1989).
36. A. Shrake and J. A. Rupley, *J. Mol. Biol.*, **79**, 351 (1973).
37. H. Wang and C. Levinthal, *J. Comp. Chem.*, **12**, 868 (1991).
38. J. L. Pascual-Ahuir and E. Silla, *J. Comp. Chem.*, **12**, 1047 (1991).
39. E. Silla, I. Tunon, and J. L. Pascual-Ahuir, *J. Comp. Chem.*, **12**, 1077 (1991).
40. S. M. LeGrand and K. M. M. Merz, Jr., *J. Comp. Chem.*, **14**, 349 (1993).
41. R. A. Abagyan, M. M. Totrov, and D. Kuznetsov, *J. Comp Chem.*, **15**, 488 (1994).
42. R. Sedgewick, *Algorithms in C++* (2nd ed.), Addison-Wesley, Reading, MA, 1992, pp. 112–114 and pp. 373–386.
43. W. F. van Gunsteren, H. J. C. Berendsen, F. Colonna, D Perahia, J. P. Hollenberg, and D. Lelouch, *J. Comp. Chem.*, **5**, 272 (1984).
44. V. Yip and R. Elber, *J. Comp. Chem.*, **10**, 921 (1989).
45. D. B. Beglov and A. A. Lipanov, *J. Biomol. Struct. Dyn.*, **9**, 205 (1991).
46. W. Kabsch and C. Sander, *Biopolymers*, **22**, 2577 (1983).
47. M. L. Connolly, *J. Am. Chem. Soc.*, **107**, 1118 (1985).
48. H. S. M. Coxeter, In *A Spectrum of Mathematics*, J. C Butcher, Ed., Auckland University Press and Oxford University Press, 1972, pp. 98–107.
49. H. Wenninger, *Spherical Models*, Cambridge University Press, Cambridge, UK, 1979.
50. C. J. Kitrick, *Structural Topology*, **11**, 15 (1985).
51. C. J. Kitrick, *Int. J. Space Struct.*, Special Issue on Geodesic Forms, **5**, 223 (1990).
52. F. P. Preparata and M. I. Shamos, *Computational Geometry. An Introduction*, Springer-Verlag, New York, 1985.
53. F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, *J. Mol. Biol.*, **112**, 535 (1977).
54. E. E. Abola, F. C. Bernstein, S. H. Bryant, T. F. Koetzle, and J. Weng, In *Crystallographic Databases—Information Content, Software Systems and Scientific Applications*, F. H. Allen, G Berghoff, and R. Sivers, Eds., International Union of Crystallography, Bonn/Cambridge/Chester, 1987, p. 107.
55. P. Alard and S. J. Wodak, *J. Comp. Chem.*, **12**, 918 (1991).

3

DETECTING HYDROPHOBIC PATCHES

P. Lijnzaad, H.J.C. Berendsen, and P. Argos.
A method for detecting hydrophobic patches on protein surfaces.
Proteins **26**:192–203 (1996)

A Method for Detecting Hydrophobic Patches on Protein Surfaces

Philip Lijnzaad,[1] Herman J.C. Berendsen,[2] and Patrick Argos[1]
[1]European Molecular Biology Laboratory, 69012 Heidelberg, Germany; [2]Department of Physical Chemistry, University of Groningen, 9747 AG Groningen, The Netherlands

ABSTRACT **A method for the detection of hydrophobic patches on the surfaces of protein tertiary structures is presented. It delineates explicit contiguous pieces of surface of arbitrary size and shape that consist solely of carbon and sulphur atoms using a dot representation of the solvent-accessible surface. The technique is also useful in detecting surface segments with other characteristics, such as polar patches. Its potential as a tool in the study of protein–protein interactions and substrate recognition is demonstrated by applying the method to myoglobin, Leu/Ile/Val-binding protein, lipase, lysozyme, azurin, triose phosphate isomerase, carbonic anhydrase, and phosphoglycerate kinase. Only the largest patches, having sizes exceeding random expectation, are deemed meaningful. In addition to well-known hydrophobic patches on these proteins, a number of other patches are found, and their significance is discussed. The method is simple, fast, and robust. The program text is obtainable by anonymous ftp.**

Key words: molecular recognition, myoglobin, Leu, Ile, Val binding protein, lipase, lysozyme, azurin, triose phosphate isomerase, carbonic anhydrase, phosphoglycerate kinase

INTRODUCTION

Biological processes such as catalysis, regulation, and protein-DNA interaction all involve acts of molecular recognition, governed by an interaction between surface portions of the partner molecules that are complementary in steric, electrostatic, and hydrophobic aspects. The latter often assume the form of a contact between hydrophobic patches, loosely defined as regions of high hydrophobicity. The origin of this interaction, the hydrophobic effect, is the large cohesion of water that results from its hydrogen bonding capabilities, and it holds sway as one of the principal driving forces of protein folding[1,2] and of protein–protein association.[3–5]

The exposure to solvent of large hydrophobic patches is unfavorable, yet they are frequently found on proteins. Their presence can often be linked to a biological role. Numerous cases are known where binding of substrates and cofactors to an enzyme involves interaction with exposed hydrophobic area, frequently in surface clefts. A number of election transfer proteins interact with their cognate proteins via hydrophobic patches.[6–9] Hydrophobic patches in the interface region of subunits are in many cases essential for oligomerization such that their alteration drastically affects the association between subunits.[10–14] Protein–DNA recognition also depends on hydrophobic interaction, for instance, in bacteriophage 434 repressor. Here, a hydrophobic pocket formed by Thr and Arg side chains interacts with a thymidine methyl group upon association.[15]

Hydrophobic patches are clearly pertinent to the study of protein function, but to our knowledge, few methods exist for detecting them automatically in a given three-dimensional structure. Korn and Burnett[16] describe a measure for surface hydrophobicity and a method to assess the hydrophobic complementarity in subunits. Their method is not suited to the analysis of patches but deals with larger interface surfaces. A different venue is taken in the use of hydrophobicity (or lipophilicity) potentials[17,18] where the hydrophobic potential at a point in space in the vicinity of the molecule is given by the sum of hydrophobicities from nearby atoms or groups, weighted by a function decreasing with distance. The potential is visualized by iso-potential surfaces or, more often, by coloring the points on the molecular surface according to their potential. In this methodology, a hydrophobic patch is the area on the molecular surface that has an associated potential exceeding a certain threshold. Although they have proved useful for comparative studies,[19,20] the physical meaning of such hydrophobic potentials is unclear. Young and co-workers[21] describe a similar approach, summing hydrophobicity parameters of residues within a certain distance from a solvent-

Abbreviations: PDB, Protein Data Bank; LIVBP, Leu/Ile/Val-binding protein; TIM, triose phosphate isomerase; PGK, phosphoglycerate kinase; AMS, 3-acetoxymercuri-4-aminobenzenesulfonamide; NAM, N-acetyl muramic acid

Received July 18, 1995; revision accepted January 22, 1996.
Address reprint requests to Patrick Argos, European Molecular Biology Laboratory, Meyerhofstrasse 1, Postfach 10.2209, 69012 Heidelberg, Germany

accessible grid point to obtain the hydrophobicity of the cluster formed by these residues. They employ a face-centered cubic lattice to represent C_α positions and to approximate the accessible surface of the protein. Although their technique highlights regions of higher average hydrophobicity, it does not delineate patches as the size of the residue clusters is determined by the lattice spacing and the radius used to identify them. Moreover, the clusters may overlap.

Here we report on a technique for identifying explicit hydrophobic patches consisting of exposed groups of neighboring apolar atoms, generally bordered by hydrophilic atoms. It is conceptually simple, fast, and robust and can easily accommodate different measures of atomic similarity. To illustrate the method, we applied it to a number of proteins, each emphasizing different structural themes: lack of significant patches, patches for substrate recognition, patches for protein recognition, patches in subunit interaction, and finally patches of yet unknown function. Myoglobin, an oxygen storage protein, represents the case where hydrophobic patches do not seem important for function. Leu/Ile/Val-binding protein (LIVBP), lipase, and lysozyme have patches involved in substrate binding. Azurin serves as an example for the electron transport group. The monomer of trypanosomal triose phosphate isomerase (TIM) displays a patch necessary for subunit interaction. Finally, carbonic anhydrase (CA) and phosphoglycerate kinase (PGK) show large hydrophobic regions that were unexpected.

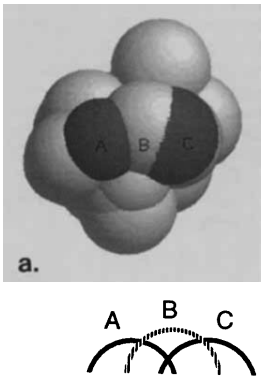

Fig. 1. Atoms A and C have overlapping extended radii. Their intersection lies beneath the accessible surface of atom B; however, they do not possess adjacent solvent-accessible surfaces. **a:** Solid model rendering. This image and the other space-filling models were prepared with the molecular display program RasMol, written by Roger Sayle of Glaxo R&D, Greenford, UK (ftp://ftp.dcs.ed.ac.uk/pub/rasmol on World Wide Web). **b:** Cross section in a plane through the centers of atom A and C.

MATERIALS AND METHODS

The accessible surface of a protein is that traced by the center of a solvent molecule ("probe") rolling over the exposed atomic surface, where atoms are considered spheres of appropriate van der Waals radius. The accessible area is proportional to the number of water molecules that can be in contact with exposed protein atoms.[22] It is obtained by incrementing the atomic radii with the probe radius and calculating the surface of the resulting set of intersecting spheres. A patch can now be defined as a contiguous piece of solvent-accessible surface made up solely of a cluster of neighboring atoms that fulfil a certain condition. For hydrophobic patches, we will use the straightforward condition that the surface atoms be apolar, i.e., they should be either carbon or sulphur. For ease of discussion, we will restrict ourselves to this classification, but the use of other, more sophisticated measures of atomic similarity is entirely possible.

Our main concern is not simply the contact or intersection of atoms but the "adjacency" of their solvent-accessible parts, i.e., having contiguous surfaces. If an atom lies within the extended radius of another atom, their solvent-accessible surfaces need not be adjacent, as shown in Figure 1. An explicit analytical description in terms of accessible arcs of intersection circles is needed to ascertain adjacency.[23] This is computationally demanding, and the topology of arcs is not resolved satisfactorily. Therefore, we resort to a surface represented by dots as obtained with the extremely fast DCLM algorithm of Eisenhaber et al.[24] Their method, based on that of Shrake and Rupley,[25] consists of superimposing regular spherical point distributions onto the atom centers and deleting all points occluded by other atoms. For the point distribution, the vertices of optimized regular polyhedra are used. Adjacency of atoms can now be defined in terms of the planar triangulation of the points thus obtained. If the vertices of a triangle lie on two or three different atoms, then this triangle's edges connect two or three different atoms and thus establish their adjacency. There is, however, no need for a full triangulation since localizing these interatomic connections is sufficient.

The inter-atom connections in the triangulation are between points lying on the boundaries of solvent-accessible portions of any one atom. These boundary points are identified by the fact that some or all of their neighboring points (in the regular polyhedron representing the atom) are buried by neighboring atoms. Boundary points on two intersecting atoms may form an edge in a surface triangulation if they are separated no more than a threshold distance δ, which depends on the distance

between the two neighboring atoms, their radii, and the point densities of the spheres representing them.

The threshold distance can be estimated by establishing the configuration of the two polyhedra that leads to a maximum in the distance between nearest points on two adjacent atoms. There are two such configurations: I (Fig. 2a) and II (Fig. 2b). The maximum of the distances (the lengths of the line segments pq in Fig. 2) in both cases is the threshold distance δ. Considering the faces of the polyhedra equilateral triangles, we find (Fig. 2a,b)

$$\delta^2 = a^2 + \frac{1}{4}b^2 \quad \text{in configuration I}$$

and

$$\delta^2 = a^2 + b^2 - \sqrt{3}ab\cos(\phi) \quad \text{in configuration II.}$$

Constants a and b are the largest and smallest lengths, respectively, of edges in the regular polyhedra (line segments pp' and qq' in the figure). The angle ϕ ($pp'm$ in Fig. 2b; see also Fig. 2c) is that of a wedge that fits snugly between the two polyhedra. If two spheres do not overlap greatly, then this wedge angle is small and configuration I will lead to a maximum edge length; otherwise configuration II yields the maximum. For typical proteins, configuration II is selected for 95% of the atom pairs due to the generally large overlap that is the result of the covalent structure and the large radii arising from the addition of the probe radius.

The wedge angle ϕ is estimated as the sum of two angles, ψ and χ. The angle ψ (Fig. 2c) is that between planes tangent to the intersection circle and both spheres and is given by the rule of cosines. It is smaller than the wedge angle ϕ by an amount χ, which is a small positive quantity that accounts for the truncation of the two spherical surfaces by one edge and one face of either polyhedron (edge pp' and face $qq'q$ in Fig. 2b). Its value depends only on the densities of points on the two spheres and is obtained from spherical trigonometry. Thus,

$$\phi = \psi + \chi$$

$$\psi = \pi - \arccos\left(\frac{r_a^2 + r_b^2 - d_{ab}^2}{2r_a r_b}\right)$$

and

$$\chi = \frac{1}{2}\alpha + \arccos\left(\frac{\cos(\beta)}{\cos\left(\frac{1}{2}\beta\right)}\right)$$

The term d_{ab} represents the distance between the centers of the considered atomic pair, each with respective radius r_a and r_b; α and β are the average arc lengths of edges in the polyhedra representing the largest and smallest sphere, respectively. Occasionally, isolated points in cavity-like regions are found that cannot form edges shorter than δ because there

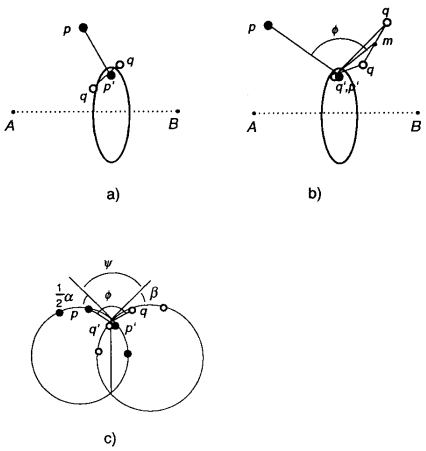

Fig. 2. Two configurations of edges of the regular polyhedra leading to the largest distance between nearest points p and q on different atoms. **a:** Configuration I. A schematic perspective view is shown, in which atom centers are represented by a small dot, the interatomic axis by a dotted line, the intersection circle by an ellipse, and relevant points and edges of the regular polyhedron by large dots and drawn lines, respectively. Point p on atom A has a maximal distance to the intersection circle; its neighboring point p' is just buried. Two equivalent points q on the neighboring atom B are both just accessible. The edge qq lies perpendicular to the edge pp' and is bisected by point p'. The distance pq is maximal since shifting the edge qq along its own axis or alternatively decreasing p's or increasing q's distance to the intersection circle will bring p and q into closer proximity. **b:** Configuration II. Symbols are as in a. The distance between p and one of the two equivalent points q is maximal if their respective neighbor points p' and q' are just buried and coincide such that both p and the point m (the midpoint of the edge qq) have a maximal distance to the intersection circle. The edge pp' bisects the angle $qq'q$. **c:** Configuration II. Explanation of symbols used in the text. Side view: atoms are shown as circles and the intersection circle's projection as a line.

are no solvent-accessible points in the vicinity. They are very few in number and of little consequence since they do not represent the main surface.

Having localized among all intersecting pairs of solvent-accessible atoms all the edges of length shorter than the threshold distance δ, patches are delineated by a simple flooding algorithm, consisting of a depth-first search started at an unmarked apolar atom. During the search, unmarked apolar atoms are marked and incorporated into the patch if they have an edge to an atom already in the patch. Their recognition is complete if no further connected atoms can be found; a new unmarked atom is searched in the linear sequence of atoms, and the process is reiterated. The process terminates when no unmarked atoms remain. An atom may have several separate pieces of exposed surface that may contribute to the same patch or to different patches.

Although the surface of a patch is, by our definition, fully connected, it may contain "gaps" formed

by hydrophilic atoms. In fact, for most protein structures, one very large and contiguous hydrophobic segment, spanning the entire protein surface and dotted by hydrophilic "islands," is found. In other words, the hydrophobic surface "percolates" in a two-dimensional sense.[26] The reason for this is the high fraction of hydrophobic atoms on the surface (48–65% in a test set of 112 unrelated monomeric proteins). The contiguous hydrophobic surface piece will consist of larger, more open regions of clusters of apolar atoms, connected by smaller "passages" or channels of hydrophobic surface. To obtain the open regions the radii of the exposed polar atoms can be expanded by a given amount (in addition to the usual extension by a probe radius yielding the accessible surface). In this way, polar atoms will become more easily adjacent, fencing off hydrophobic patches and closing the corridors connecting the open regions. A physical justification for this lies in the dominance of the polar interaction of a solvent molecule that is in contact with both a hydrophilic and a hydrophobic atom. This hydration effect can be modeled by increasing the radius of the polar atom.

The exact increment for the polar radii cannot be deduced from first principles. Nonetheless, it should relate to the size of the solvent probe. In this work, we use a polar expansion equal to the radius of the probe, 1.4 Å. There are two reasons for this choice. First, two polar atoms, separated by their atomic radii plus four probe radii, can be exactly bridged by two water molecules, providing a natural boundary for a hydrophobic surface segment (Fig. 3). Thus, the expanded polar atom represents the solvated polar atom in an approximate way. Second, the patches so defined and refined, as explained below, correspond well to the intuitive notion of a hydrophobic patch when inspected graphically. Interestingly, the average ratio of the number of hydrophobic patches relative to the number of hydrophilic patches in each protein has a maximum for an expansion radius of 1.4 Å, implying a maximal contrast and facilitating the automated pattern recognition (Fig. 4). The patches generally contain no gaps, but an isolated polar group in the middle of a large hydrophobic region (e.g., the oxygen atom of a tyrosine) is allowed since it does not disrupt the contiguity of the patch surface.

Although the central kernel of a hydrophobic patch is well defined employing this scheme, a considerable amount of hydrophobic surface area at the border of patches can be buried by the polar expansion. To circumvent this, exposed apolar atoms covered by the polar expansion and directly adjacent to an atom in the preliminary patch are added to the preliminary patch. If such atoms are adjacent to several preliminary patches, they are added to the largest. The recovery of buried atoms results in a total hydrophobic patch area that is considerably larger

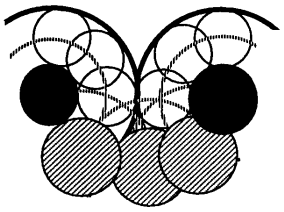

Fig. 3. Two polar atoms separated by four solvent radii. Black circles indicate polar atoms; shaded circles, apolar atoms; open circles, possible locations of hydration waters; dashed open circles, solvent-accessible surfaces of atoms; thick open circles, expanded polar atoms. The two hydration waters that are in contact with both a polar and an apolar atom touch and form a natural boundary, dividing the hydrophobic surface they cover.

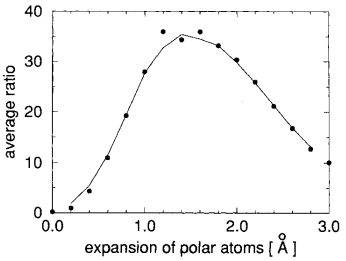

Fig. 4. The average ratio of the number of hydrophobic patches relative to the number of hydrophilic patches as a function of the polar expansion, observed in 112 non-homologous monomeric proteins. The curve represents a running average over three data points (sampled at 0.2 Å intervals) and shows a maximum at an expansion of 1.4 Å.

than that of the preliminary patches of size larger than 10 Å^2. However, not all apolar area is recovered because the polar expansion covers more atoms than are directly adjacent to the preliminary patch. This concerns mostly area in passages connecting preliminary patches that is deliberately removed. Further, some small patches are completely buried by the polar expansion before the refinement and thus are never identified as preliminary patches. About one-fifth of the total apolar area is not assigned to hydrophobic patches on average, as a result of the above two effects.

Randomization

To contrast the results from the patch calculation with statistical expectation, it is necessary to randomize the surface of the protein prior to the patch calculation. A realistic shuffling of the surface would involve a random replacement of side chains, followed by an energy minimization to remove unfavorable interactions, which is a complex and time-

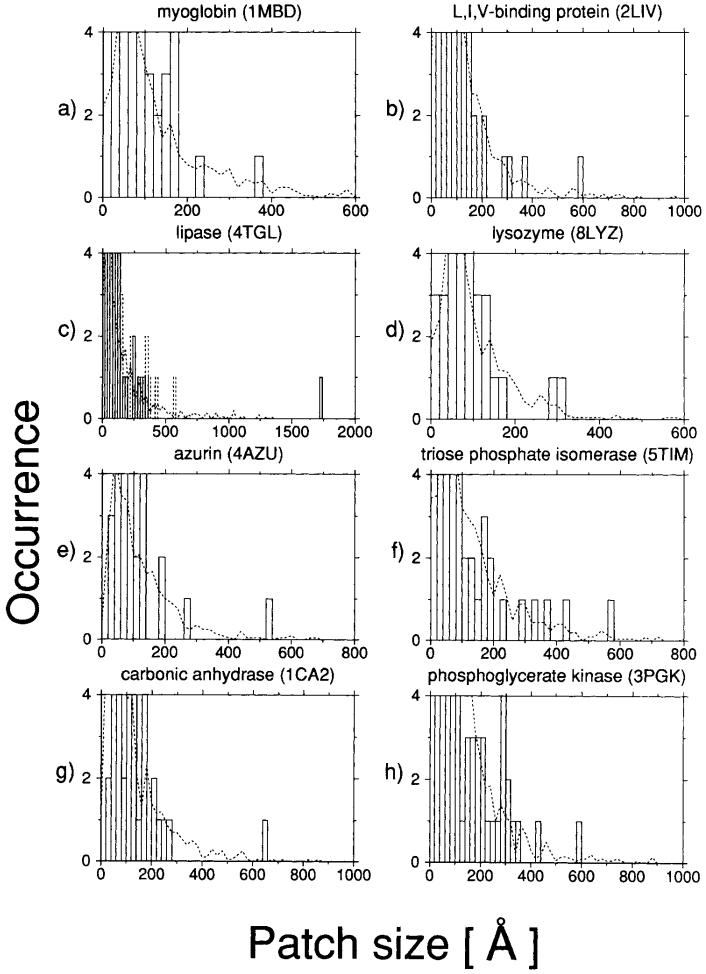

Fig. 5. **a–h:** Occurrence of patches per accessible area size class (bin size 10 Å^2), shown as histograms. Plotted horizontally is the size class; the height of the bars represents the number of observed patches while the dotted line indicates the mean number of patches in the bin averaged over 20 randomized structures. For lipase, the dotted bars represent the patches found in the closed form (PDB code 3TGL).

consuming process. We thus adopted the following procedure. The total oxygen and nitrogen area was calculated; all atoms were then assigned as carbon atoms. Next, an atom with solvent accessibility greater than 10 Å^2 was selected randomly and designated as an oxygen atom. This cut-off was chosen since it is roughly equal to the protein-accessible area covered by one water molecule of a solvent monolayer over the entire surface.[27] The selection process was repeated until the sum of the areas of the selected atoms equalled the total oxygen area initially calculated; the same was effected for the nitrogen atoms. The constraints on the vicinity of polar and apolar atoms implied by the covalent

structure are thus removed, whereas the principal determinants of the protein's patch size distribution, namely, hydrophobic surface fraction and surface geometry, are conserved.

Data

High-resolution X-ray structures of the proteins discussed here were obtained from the Brookhaven Protein Data Bank (PDB).[28] Cofactors and prosthetic groups were included in the calculations; solvent molecules, ions, and hydrogen atoms, if present, were excluded. Atomic radii employed were those of Eisenberg and McLachlan[29]: N, 1.7 Å; O, 1.4 Å; C and S, 1.9 Å; the radius of the solvent probe was set to 1.4 Å[22,23], as was the polar expansion. Randomizations were performed 20 times for each protein. In the examples discussed, 252 points per atom were used. Total execution times for a given protein were in the order of seconds (DECstation 3000, Alpha processor operating at 125 MHz), due to the very high computational efficiency of DCLM.[24]

The input to the patch-delineating program consists of a protein structure in PDB format; the user is notified of missing atoms and alternate side chain conformations that may affect the results. Atomic radii, probe radius, polar expansion, and point density per atom can be specified, the latter currently restricted to numbers between 12 and 252. The surface can optionally be randomized prior to the calculation. The program produces a list of surface patches, sorted by size. For each patch, the accessible area, number of atoms, number of surface points, and identities and surface areas of all constituent atoms are given. The program is written in ANSI C and runs under a variety of UNIX systems: Ultrix, OSF1, SunOS, and Irix. The source code is available from http://www.embl-heidelberg.de/argos/quilt/quilt_info.html on World Wide Web.

RESULTS AND DISCUSSION

Figure 5 shows the distribution of patch sizes, together with the results from randomization. Details concerning sizes and atomic composition of the largest patches are given in Table I. The patches in each protein are ranked by size, with the largest patch numbered 0, the second largest 1, and so forth. Two examples of the position and extent of the patches on the protein surface are depicted in Figure 6. Results for individual proteins will be addressed below.

We assume that only large patches are biologically interesting. Statistically, the likelihood of large patches should be less than that of smaller patches, and this is seen in the histograms of observed and randomly generated patches (Fig. 5). However, even rather large patches will occur by chance alone. Only when the occurrence of a patch of a certain size is at odds with the likelihood of its occurrence in the random case does a patch seem material and possibly purposeful.

A problem arises when more than one patch is observed in a particular size class. If the boundaries of the size classes are chosen differently, such patches could fall into different ranges, whereas the base line will remain the same. The height of the new peaks can now again be considered an indicator of the significance of a peak. However, this situation does not occur in the examples described below.

As shown in Figure 5, only the largest few patches in each protein are well above the base line, and their occurrence is clearly separated from the main distribution. These patches generally coincide with regions of known biological relevance (see below). This lends confidence to the use of randomization as a rough guide for the significance of patches. The criterion is stringent since lifting the constraints on the covalent structure should allow for larger patches.

The area of the significant patches is generally larger than 300 Å^2. The free energy associated with the solvation of apolar surface is generally believed to be proportional to its area (see Eisenberg and McLachlan[29] and references therein). Using 30 cal/molÅ^2 as an estimate,[30] the energetic cost of a patch of 300 Å^2 would be around 9 kcal/mol. This is substantial, considering that the free energy of one hydrogen bond has been estimated around 2 *kcal/mol*[31] and that of folding at the same order of magnitude.[32] This suggests that large hydrophobic patches are there by design and do not represent coincidences tolerated by the protein.

Although 300 Å^2 is roughly one and a half times the accessible area of an average, fully exposed hydrophobic residue according to Rose et al.,[33] the patches are in fact composed of several residues. The main constituents are hydrophobic residues, but other amino acid types frequently contribute apolar atoms. This can be significant, as exemplified by the Thr/Arg patch of bacteriophage 434 repressor referred to in the Introduction. The general trends regarding composition, size, and other characteristics will be addressed in the accompanying paper.

Polar Expansion Radius

The largest and most biologically significant patches of lysozyme, azurin, and trypanosomal TIM were used to monitor the influence of the polar expansion. The results are given in Table II. The patches obtained with different choices of the expansion are depicted in Figure 7. For polar expansions larger than 0.8 Å, the changes in patch area are gradual for most of the patches, and they do not split into separate patches. This is not true for the TIM patch, where the increments in patch size level off only at 1.4 Å. These results, together with the considerations in the Methods section, suggest that 1.4 Å is a reasonable value and is therefore used throughout this work. This parameter can also be utilized to adjust the level of detail desired: a large

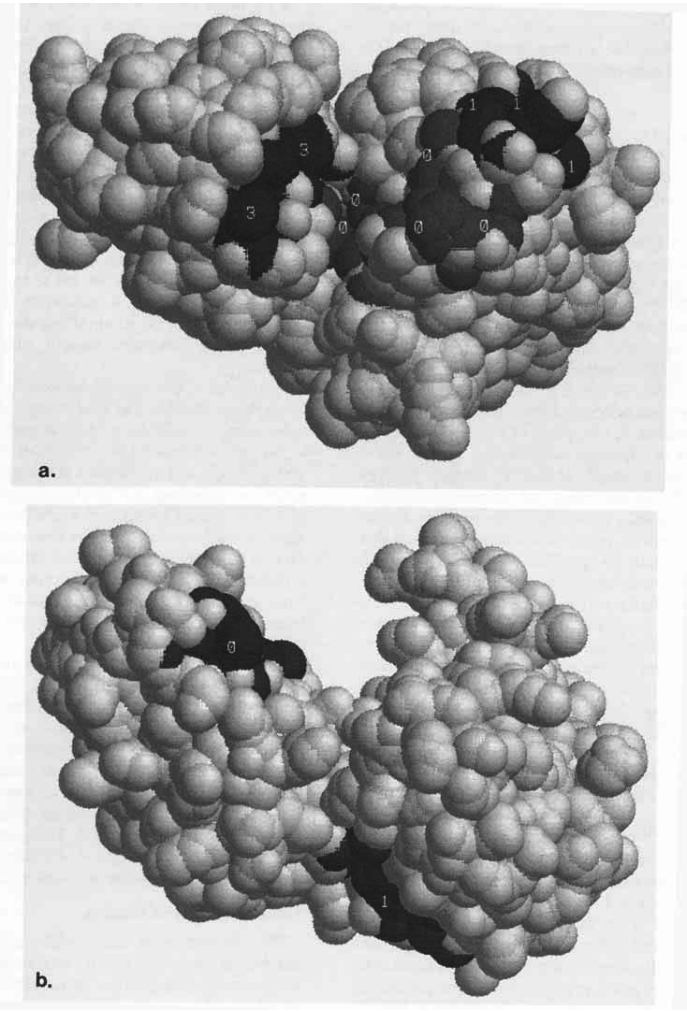

Fig. 6. Patches indicated by darker shading on the solvent-accessible surface. Rank numbers according to area are indicated as 0 (light gray), 1 (dark gray), and 3 (dark gray). **a:** Leu/Ile/Val-binding protein. The substrate binding site coincides with patch 0 (largest patch) and lies close to the hinge region between the two domains. The patch numbered 2 is on the other side of the protein structure. **b:** Phosphoglycerate kinase. The ATP-binding site forms part of patch 0.

polar expansion results in small but manifest patches, whereas a smaller value will yield larger, more diffuse patches by allowing smaller patches to become connected via hydrophobic passages.

Myoglobin

Hydrophobic patches in sperm whale myoglobin are unlikely to be important for function. The substrate of this monomeric protein, O_2, is small and

TABLE I. Atomic Composition of Larger Hydrophobic Patches

Protein*	Patch†	Size‡	Atoms§	Remarks
Myoglobin (1MBD)	0	366	**V1:** α, β, γ1, γ2; **L2:** δ1; **S3:** α; **W7:** ζ2; **G80:** α; **E136:** C, β; **L137:** α, β, δ1, δ2; **K140:** α, β, γ, δ, ϵ; **D141:** γ; **A143:** β; **A144:** β	Not significant
LIV-binding protein (2LIV)	0	582	**S12:** C, β; **G13:** C, α; **P14:** C, α, β, δ, γ; **V15:** γ1, γ2; **A16:** β; **Y18:** β, δ2, ϵ1, ϵ2, ζ; **A52:** β; **C53:** β; **L77:** C, δ2, γ; **C78:** α, β, SG; **S80:** β; **A100:** β; **T102:** γ2; **T118:** γ2; **G119:** C; **E226:** δ; **K248:** δ; **P249:** β, γ; **Y252:** δ2, ϵ2; **G274:** α; **A275:** α, β; **F276:** ϵ1, ϵ2, ζ; **T279:** γ2; **M313:** ϵ; **F327:** ζ; **F329:** ϵ1, ζ;	Substrate binding site
	1	372	**D54:** α, γ; **P55:** C, β, δ, γ; **K56:** α, β, δ, ϵ, γ; **Q57:** α; **V59:** β, γ1, γ2; **A60:** α, β; **P84:** α, β, δ, γ; **D87:** β; **I88:** α, δ1, γ1, γ2; **D91:** β;	
Lipase, "open" form (4TGL)	0	1739	**S83:** α; **S84:** α, β; **I85:** α, β, γ1, γ2, δ2, δ1; **R86:** α, γ, δ; **W88:** C, β, γ, δ1, δ2, ϵ2, ϵ3, ζ2, ζ3, η2; **I89:** C, β, γ1, γ2, δ1; **A90:** α; **D91:** C, β; **L92:** α, C, β, γ, δ1, δ2; **T93:** α, β, γ2; **F94:** δ1, δ2, ϵ1, ϵ2, ζ; **V95:** α, β, γ1, γ2; **P96:** β, γ, δ; **V97:** α, γ1, γ2; **K109:** γ, δ, ϵ; **G110:** α; **F111:** ζ; **L112:** δ2; **P177:** β; **I204:** γ1, γ2, δ1; **V205:** γ1, γ2; **H207:** β; **L208:** C, β, δ1, δ2; **P209:** α, γ, δ; **P210:** α, β, γ, δ; **A212:** α, C, β; **F213:** α, β, γ, δ1, δ2, ϵ1, ϵ2, ζ; **F215:** ϵ1; **T252:** C, γ2; **S253:** α, C; **V254:** β, γ1, γ2; **L255:** C, β, γ, δ1, δ2; **D256:** α; **H257:** β, δ2; **L258:** β, δ1, δ2; **N264:** β; **L267:** α, β, δ1, δ2; **T269:** C, β, γ2;	Lipid binding face
Lysozyme (8LYZ)	0	300	**R61:** δ, ζ; **W62:** β, γ, δ1, δ2, ϵ2, ϵ3, ζ2, ζ3, η2; **W63:** ζ2, ζ3, η2; **R73:** β, γ, δ; **N74:** C; **L75:** α, β, γ, δ1, δ2; **I98:** γ2; **D101:** β, γ; **A107:** β;	B and C subsites
	1	289	**N65:** β; **D66:** α; **N77:** C; **I78:** α, C, β, γ1, δ1; **P79:** α, β, γ, δ; **S81:** β; **A82:** α, β; **T89:** β, γ2; **A90:** α, β;	Unclear; conserved
Azurin, pH 5.5 (4AZU)	0	528	**D11:** β; **Q12:** α, C; **M13:** γ, SD, ϵ; **Q14:** γ; **N38:** β; **L39:** γ, δ1, δ2; **P40:** α, β, γ, δ; **K41:** β; **V43:** α, γ1, γ2; **M44:** ϵ; **G116:** α, C; **H117:** δ2; **A119:** α, β; **L120:** α, β, δ1, δ2;	Involved in electron transfer
Triose phosphate isomerase, subunit A (5TIM)	0	568	**T44:** α, β; **F45:** C, β, δ1, δ2, ϵ2; **V46:** α, β, γ1, γ2; **H47:** δ2, ϵ1; **L48:** β, δ1; **A49:** α, C, β; **M50:** β; **Q65:** β; **N66:** β; **S71:** α, β; **G76:** C; **E77:** α, C; **V78:** γ1, γ2; **S79:** β; **P81:** C, β; **I82:** α, γ1, γ2, δ1; **L83:** δ2; **F86:** α, β, γ, δ1, δ2, ϵ1, ϵ2; ζ;	Interface patch; deletion results in lack of dimerization
Carbonic anhydrase (1CA2)	0	656	**H4:** ϵ1; **W5:** δ1; **D19:** β, γ; **F20:** α, δ1, ϵ1, ϵ2, ζ; **P21:** γ, δ; **I22:** α, β, γ1, γ2, δ1; **K24:** C; **G25:** α; **E26:** γ; **E69:** β, γ; **I91:** γ1, γ2, δ1; **Q92:** γ; **H94:** ϵ1; **V121:** γ1, γ2; **W123:** ζ2; **F131:** β, δ1, ϵ1, ζ; **G132:** α; **V135:** γ1, γ2; **Q136:** α, γ; **V143:** γ1, γ2; **L198:** α, β, δ1, δ2; **P201:** C, β; **P202:** β, γ, δ; **L204:** β, δ1, δ1, δ2; **E205:** β; **W209:** ζ2;	Part of funnel leading to active site
Phosphoglycerate kinase (3PGK)	0	586	**ATP:** C1*, C2; **A212:** β; **K213:** α, β, γ, δ, ϵ; **A215:** β; **M237:** SD; **D256:** γ; **I285:** δ1; **F289:** α, β, γ, δ1, δ2, ϵ1, ϵ2, ζ; **S290:** α, β; **A291:** β; **A306:** α, C; **G307:** C; **W308:** β, γ, δ2, ϵ2, ϵ3, ζ2, ζ3, η2; **Q309:** C; **L311:** α, δ2; **V339:** γ2; **E341:** β, γ, δ; **F342:** α, C, γ, δ1, δ2, ϵ1, ϵ2, ζ; **K344:** δ;	ATP recognition
	1	428	**Q9:** β, γ; **R38:** α; **V40:** γ1; **A41:** α, β; **P44:** α, β, γ, δ; **Y48:** α, β, δ2, ϵ2, ζ; **H52:** ϵ1; **F185:** C, β, γ, δ1, δ2, ϵ1, ϵ2, ζ; **L186:** α, β, γ, δ2; **K189:** β, ϵ; **V389:** γ1; **T391:** β, γ2;	Unclear

*Protein with Protein Data Bank (PDB) code in parentheses.
†Rank of the patch with 0 as largest, 1 second largest, etc.
‡Solvent-accessible area of the patch in square Ångstrøms.
§Residue type and sequence number (bold) and atoms it contributes. For brevity, all carbon atoms except the carbonyl C are given by their Greek letter identifiers; other atoms are given by their PDB-file designations.

TABLE II. Patch Sizes as a Function of the Polar Expansion*

Patch	0.8 Å			1.0 Å			1.2 Å			1.4 Å			1.6 Å			1.8 Å		
Lysozyme	9	28	307	9	28	307	9	28	307	9	27	300	7	22	253	6	121	240
Azurin	17	37	601	15	30	530	15	30	530	14	29	528	14	29	528	14	29	528
TIM-A	21	46	509	21	45	497	18	45	581	18	44	568	18	44	568	14	30	412

*The entries in the table relating to one patch are, in order, number of residues, number of atoms and solvent-accessible area in $Å^2$. The patches refer to the largest patches in Table I.

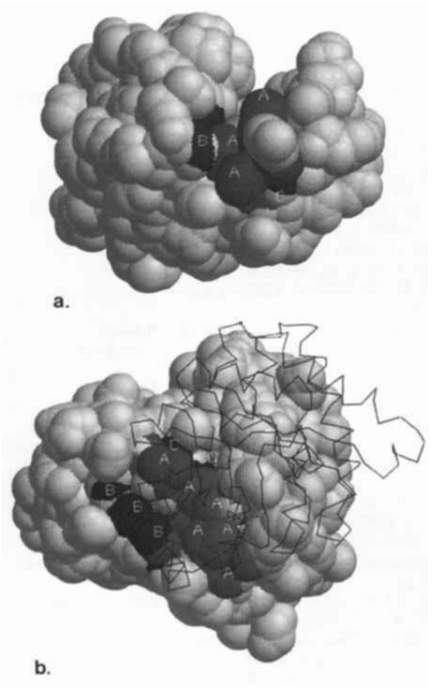

Fig. 7. Influence of the polar expansion. The patches from Table I with known biological significance are shown in a darker shade on the solvent-accessible surface. Area labeled A (light gray) designates the patch obtained with an expansion of 1.8 Å. The patch surface obtained by using 1.4 and 1.0 Å is given by A + B and A + B + C, respectively. **a:** Hen egg-white lysozyme. The patch depicted is at the entrance of the active cleft. **b:** Triosephosphate isomerase, chain A. For reference, the C_α trace of chain B is also shown; the view is roughly in the plane of the dimer interface.

ligates to the iron atom in the heme. The protein has no interactions with other proteins in vivo; it functions independently and has no cofactors or noncompetitive inhibitors. As can be seen in Figure 5a, the actual patches, even the larger ones, generally follow random expectation. In fact, the occurrence of smaller patches is above those resulting from the-random trials. This is almost always the case for the protein structures examined here. Though the largest patch has sizeable area, it does not include any portions of the heme group. Though parts of the heme contribute to the fifth largest patch (not shown), they do not form an appreciable hydrophobic surface. This shows that the heme is well shielded from the solvent and is, in hydrophobic respects, indistinguishable from the rest of the protein surface.

Leu/Ile/Val-Binding Protein

This extracellular transport protein binds aliphatic amino acids with high affinity. The binding of the hydrophobic moiety occurs in a surface cleft lined by Tyr-18, Leu-77, Cys-78, Ala-100, Ala-101, and Phe-276.[34] Substrate binding induces domain movement such that the cleft closes, rendering the bound ligand fully buried. The largest and most significant patch identified by our procedure (Figs. 5b, 6a) coincides with this binding site, even though the hydrophobic specificity pocket itself is not fully solvent accessible and represents only a small part of the patch. It is larger than that indicated by Sack and co-workers[34] and lies in an excellent position to associate, upon cleft closure, with the patch ranked fourth (illustrated in Fig. 6a and numbered as "3"). This finding suggests that hydrophobic interaction is an important factor in the closure and the latter patch is biologically relevant. Its importance as judged by our standard is relatively low, implying that our criterion is not too lenient. The patch second in area is adjacent to the largest patch and could well be important for trapping the substrate.

Lipase

Two structures of lipase from the fungus *Rhizomucor miehei* have been determined: the closed, inactive form and the open, active form.[35,26] In the active form, the so-called lid, consisting of a short helix and a stretch of random coil, is rolled back, exposing a large hydrophobic face that can interact with a fatty substrate. The resulting patch of 1,739 $Å^2$ is visible as a distinct peak in Figure 5c. In the closed, inactive form, two large patches can be distinguished (dotted bars in Fig. 5c). Together they occupy about the same region as the large patch in the open form. The larger of the two is composed of a hydrophobic surface that cannot be covered by the

lid and consists of part of the lid and of atoms from residues 203–214 and 252–254 that form the boundary of the patch in the active, open form (not shown). The purpose of this remaining hydrophobic patch is unknown. It could be a result of the difficulty in constructing a lid that completely covers the active cleft, or alternatively it could enable activation by providing a recognition site such that the substrate can prise the lid open.

Lysozyme

Most of the well-known substate binding interactions in hen egg white lysozyme concern hydrogen bonds.[37] Still, some hydrophobic interactions are important; for instance, Trp-62 makes an extensive contact with the ring of the N-acetyl muramic acid (NAM) unit of the bound polysaccharide chain[38] and Trp-63, also exposed, determines cleavage specificity.[37] The largest hydrophobic patch found by our method comprises two sub-sites of the polysaccharide binding cleft that contain the tryptophans (Fig. 5d; see also Fig. 7a, where area A + B delineates the patch). The other patch is at the opposite side of the molecule, centered around the exposed side chain of Ile 78. Its presence is puzzling as this site is not involved in substrate binding or catalysis. The residues constituting it, particularly the hydrophobic character of position 78, are conserved in the alignment given by 3DALI, a database that documents the alignment of homologous sequences from proteins of known tertiary structure through spatial equivalencing.[39] The conservation supports the biological significance of the lysozyme patch, which is likely important for the in vivo function of the enzyme. An interesting case is the seventh largest patch; it coincides with the binding site of bromophenol red in the vicinity of the F sub-site as found by Kadhusudan and Vijayan.[40] Binding of this dye affects the activity of the enzyme in vivo, but not in vitro; the authors speculate on its role in docking onto the polysaccharide fiber. This example again points to the stringency of our significance threshold and illustrates the relevance of our method to drug docking studies by suggesting binding sites for hydrophobic compounds.

Azurin

Recently, the involvement in electron transfer of an azurin hydrophobic patch has been demonstrated experimentally.[7] Through site-directed mutagenesis it was found that the patch surrounding the conserved Met-44 is engaged in reactions with both cytochrome c_{551} and nitrite reductase. Met-44 forms the center of the largest patch identified by our method, which is well above the significance threshold (Fig. 5e). All the hydrophobic amino acids in this patch are conserved in the 3DALI alignment.

Triose Phosphate Isomerase Subunit

As an example of the role hydrophobic patches play in subunit interactions, we studied the subunit of TIM, which is a homodimer. A large hydrophobic patch forms part of the interface of trypanosomal TIM. A mutant designed to disrupt this patch proved to be an active and stable monomer,[41] demonstrating the importance of the patch in dimerization. Our algorithm finds it the largest in TIM (Fig. 5f; see also Fig. 7b). One other large patch could be significant. TIM from *Escherichia coli,* yeast, and humans, but not chicken, also has patches in roughly the same spatial position (data not shown). Presently, no biological relevance can be assigned to them.

Carbonic Anhydrase II

The enzyme carbonic anhydrase II (CA) catalyzes the reaction $CO_2 + H_2O \leftrightarrow HCO_3^- + H^+$ and is one of the fastest known. Its active site consists of a zinc atom at the bottom of a funnel. The largest hydrophobic patch in human CA, which is the only significant patch noted by our procedure (Fig. 5g), forms one side of the funnel. The presence of a patch of this size (656 Å^2) is surprising. Eriksson et al.[42] have implicated this apolar region in CO_2 binding, based on crystallographic studies of complexes involving the inhibitors 3-acetoxymercuri-4-aminobenzenesulfonamide (AMS) and Diamox. The hydrophobic moieties of these compounds pack against the atoms composing the patch. However, the patch we observe is much larger than would be needed simply to bind a CO_2 molecule. It is also larger than the region found by Eriksson and co-workers,[42] who indicate Val-121, Phe-131, Val-135, Val-143, Leu-198, Val-207, and Trp-209 as constituents. Val-207 is absent in our patch due to the use of the accessible rather than the molecular surface and probably also due to the polar expansion. Most of the residues, or their hydrophobic character, are conserved as judged from the 3DALI alignment. We suggest that structurally non-disruptive mutations into more hydrophilic residues will lead to a decrease in the catalytic rate, from the following considerations. First, the trapping of CO_2 may be enhanced by the greater diffusional cross section of the binding site. Second, in the proposed catalytic mechanism, various water molecules are involved. If the funnel were too hydrophilic, their mobility could be hampered by the hydration of the hydrophilic groups. Thus, the hydrophobic "wall" provided by the largest patch could also facilitate the two-dimensional diffusion of the substrate to and from the active center.

Phosphoglycerate Kinase

The largest and most significant patch of this glycolytic enzyme is comprised by atoms of the ATP-binding site that surround the adenine (Fig. 5h; see

also Fig. 6b). Phe-289 and Trp-308 form a lid that is in contact with the adenine group. However, the hydrophobic patch is composed of more residues than these highly conserved two.[43] Phe-342, for instance, is rather remote from the ATP, yet it is part of this patch and is also conserved in the 3DALI alignment. We speculate that this patch plays a role in trapping the ATP substrate. The second largest apolar cluster lies on the side of the phosphoglycerate binding lobe, more or less perpendicular to the axis of the hinge. It is not involved in the binding of phosphoglycerate or ATP and seems unlikely to play a role in the hinge motion of the two domains. The most important contributors to this patch, Phe-185 and Leu-186, are again conserved.

The above case studies all made use of crystallographically determined protein structures, treated as immobile, rigid entities. This is a simplification as proteins are dynamic, and the internal motion is often important for function. The application of our method to several instantaneous structures along a simulated molecular dynamics trajectory is feasible, owing to the high speed of the procedure. This should shed light on many of the dynamic aspects but falls outside the scope of the current research. It is anticipated, however, that the larger patches will, on average, not change dramatically since average atomic displacements are of the order of Ångstrøms, and it seems unlikely that collective and correlated atomic movement would drastically alter a patch of 300 square Ångstrøms.

CONCLUSIONS

The method presented in this work is a helpful tool for delineating hydrophobic patches in proteins. The method is simple, fast, robust, and easily extended and requires only one adjustable parameter. Generally, only the largest patches are biologically relevant and can be rationalized in terms of protein function and interaction. This is true for Leu/Ile/Val-binding protein, active lipase, lysozyme, and carbonic anhydrase (substrate binding); azurin (electron transfer); and TIM (dimerization). In some cases the patches seem too large to be completely explained by their function. The inactive, closed form of lipase displays large patches that may be important for activation. Carbonic anhydrase has an apolar cluster of 656 $Å^2$, which seems too large for simple binding of CO_2. There remains a group of patches whose relevance cannot be explained easily. Our procedure should be useful in suggesting critical areas for further theoretical and experimental investigation such as drug docking, protein–protein interactions, dynamic behavior, and the design and engineering of binding sites.

ACKNOWLEDGMENTS

The authors thank Jaap Heringa, Frank Eisenhaber, and André Juffer for fruitful discussions. We also thank Simon Hubbard for a critical reading of the manuscript.

REFERENCES

1. Kauzmann, W. Some factors in the interpretation of protein denaturation. Adv. Protein Chem. 14:1–63, 1959.
2. Dill, K.A. Dominant forces in protein folding. Biochemistry 29:7133–7155, 1990.
3. Chothia, C., Janin, J. Principles of protein–protein recognition. Nature 256:705–708, 1975.
4. Argos, P. An investigation of protein subunit and domain interfaces. Prot. Eng. 2:101–113, 1988.
5. Janin, J., Chothia, C. The structure of protein–protein recognition sites. J. Biol. Chem. 26:16027–16030, 1990.
6. Guss, J.M., Freeman, H.C. Structure of oxidized poplar plastoycynanin at 1.6 Å resolution. J. Mol. Biol. 169:521, 1983.
7. vam de Kamp, M., Silvestrini, M.C., Brunori, M., van Beumen, J., Hali, F.C., Canters, G.W. Involvement of the hydrophobic patch of azurin in the electron-transfer reactions with cytochrome c551 and nitrite reductase. Eur. J. Biochem. 194:109–118, 1990.
8. Pelletier, H., Kraut, J. Crystal structure of a complex between electron transfer partners, cytochrome c peroxidase and cytochrome c. Science 258:1748–1755, 1992.
9. Chen, L., Durley, R.C.E., Mathews, F.S., Davidson, V.L. Structure of an electron transfer complex: Methylamine dehydrogenase, amicyanin and cytochrome c551i. Science 264:86–89, 1994.
10. Jones, D.H., McMillan, A.J., Fersht, A.R. Reversible dissociation of dimeric tyrosyl–tRNA synthetase by mutagenesis at the subunit interface. Biochemistry 24:5852–5857, 1985.
11. Ponder, J.W., Richards, F.M. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. J. Mol. Biol. 193:775–791, 1987.
12. Brange, J., Ribel, U., Hansen, J.F., Dodson, G., Hansen, M.T., Havelund, S., Melberg, S.G., Norris, F., Snel, L., Sörensen, A.G., Vogt, H.O. Monomeric insulins obtained by protein engineering and their medical implications. Nature 333:679–682, 1988.
13. Mossing, M.C., Sauer, R.T. Stable, monomeric variants of lamda-Cro obtained by insertion of a designed beta-hairpin sequence. Science 250:1712–1715, 1990.
14. Borchert, T.V., Abagyan, R., Radha Kishan, K.V., Zeelen, J.P., Wierenga, R.K. The crystal structure of an engineered monomeric triosephosphate isomerase, monoTIM: The correct modelling of an eight-residue loop. Structure 1:2105–2123, 1993.
15. Anderson, J.E., Ptashne, M., Harrison, S.C. Structure of the repressor–operator complex of bacteriophage 434. Nature 326:846–852, 1987.
16. Korn, A.P., Burnett, R.M. Distribution and complementarity of hydropathy in multisubunit proteins. Proteins 9:37–55, 1991.
17. Furet, P., Sele, A., Cohen, N.C. 3D molecular lipophilicity potential profiles: A new tool in molecular modeling. J. Mol. Graph. 6:182–189, 1988.
18. Fauchère, J.L., Quarendon, P., Kaetterer, L. Estimating and representing hydrophobicity potential. J. Mol. Graph. 6:203–207, 1988.
19. Brasseur, R. Differentiation of lipid-associating helices by use of three-dimensional molecular hydrophobicity potential calculations. J. Biol. Chem. 266:16120–16127, 1991.
20. Kellog, G.E., Semus, S.F., and Abraham, D.J. HINT: A new method of empirical hydrophobic field calculation for CoMFA. J. Comput. Aid. Des. 5:545–552, 1991.
21. Young, L., Jernigan, R.L., Covell, D.G. A role for surface hydrophobicity in protein–protein recognition. Protein Sci. 3:717–729, 1994.
22. Lee, B., Richards, F.M. The interpretation of protein structure: Estimation of static accessibility. J. Mol. Biol. 119: 537–555, 1971.
23. Connolly, M.L. Analytical surface calculation. J. Appl. Crystallogr. 16:548–558, 1983.
24. Eisenhaber, F., Lijnzaad, P., Argos, P., Sander, C., Scharf, M. The double cubic lattice method: Efficient approaches to numerical integration of surface area and volume and to

dot surface contouring of molecular assemblies. J. Comp. Chem. 16:273–284, 1995.
25. Shrake, A., Rupley, J.A. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. J. Mol. Biol. 79:351–371, 1973.
26. Feder, J. Fractals. New York: Plenum, 1988:104–148.
27. Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
28. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F.J., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. A computer based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
29. Eisenberg, D., McLachlan, A.D. Solvation energy in protein folding and binding. Nature 319:199–203, 1986.
30. Hermann, R.B. Use of solvent cavity area and number of packed solvent molecules around a solute in regard to hydrocarbon solubilities and hydrophobic interactions. Proc. Natl. Acad. Sci. USA 74:4144–4145, 1977.
31. Fersht, A.R., Shi, J.P., Knill-Jones, J., Lowe, D.M., Wilkinson, A.J., Blow, D.M., Brick, P., Carter, P., Waye, M.M.Y., Winter, G. Hydrogen bonding and biological specificity analysed by protein engineering. Nature 314:235–238, 1985.
32. Creighton, T.E. Proteins. Structures and Molecular Properties. New York: Freeman, 1984.
33. Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H., Zehfus, M.H. Hydrophobicity of amino acid residues in globular proteins. Science 229:834–838, 1985.
34. Sack, J.S., Saper, M.A., Quiocho, F.A. Periplasmic binding protein structure and function. Refined X-ray structures of the leucine/isoleucine/valine-binding protein and its complex with leucine. J. Mol. Biol. 206:171–191, 1989.
35. Brzozowski, A.M., Derewenda, U., Derewenda, Z.S., Dodson, G.G., Lawson, D.M., Turkenburg, J.P., Bjorlink, F., Huge-Jensen, B., Patkar, S.A., Thim, L. A model for interfacial activation in lipases form the structure of a fungal lipase–inhibitor complex. Nature 351:491–494, 1991.
36. Derewenda, U., Brzozowski, A.M., Lawson, D.M., Derewenda, Z.S. Catalysis at the interface: The anatomy of a conformational change in triglyceride lipase. Biochemistry 31:1532–1541, 1992.
37. Imoto, T., Johnson, L.N., Norht, A.C.T., Phillips, D.C., Rupley, J.A. 21. Vertebrate Lysozymes. New York: Academic Press, 1972:665–868.
38. Strynadka, N.C., James, M.N. Lysozyme revisited: Crystallographic evidence for distortion of an N-acetylmuramic acid residue bound in site D. J. Mol. Biol. 220:401–424, 1991.
39. Pascarella, S., Argos, P. A data bank merging related protein structures and sequences. Protein Eng. 5:121–137, 1992.
40. Kadhusudan, M., Vijayan, M. Additional binding sites in lysozyme. X-ray analysis of lysozyme complexes with bromophenol red and bromophenol blue. Protein Eng. 5:399–404, 1992.
41. Borchert, T.V., Abagyan, R., Jaenicke, R., Wierenga, R.K. Design, creation, and characterization of a stable, monomeric triosephosphate isomerase. Proc. Natl. Acad. Sci. USA 91:1515–1518, 1994.
42. Eriksson, A.E., Jones, T.A., Liljas, A. Refined structure of human carbonic anhydrase II at 2 Å resolution. Proteins 4:274–282, 1988.
43. Watson, H.C., Walker, N.P.C., Shaw, P.J., Bryant, T.N., Wendell, P.L., Fothergill, L.A., Perkins, R.E., Conroy, S.C., Dobson, M.J., Tuite, M.F., Kingsman, A.J., Kingsman, S.M. Sequence and structure of yeast phosphoglycerate kinase. EMBO J 1:1635–1640, 1982.

4

HYDROPHOBIC PATCHES

P. Lijnzaad, H.J.C. Berendsen, and P. Argos.
Hydrophobic patches on the surfaces of protein structures.
Proteins **25**:389–397 (1996)

Hydrophobic Patches on the Surfaces of Protein Structures

Philip Lijnzaad,[1] Herman J.C. Berendsen,[2] and Patrick Argos[1]
[1]European Molecular Biology Laboratory, 69012 Heidelberg, Germany; [2]Department of Physical Chemistry, University of Groningen, 9747 AG Groningen, The Netherlands

ABSTRACT **A survey of hydrophobic patches on the surface of 112 soluble, monomeric proteins is presented. The largest patch on each individual protein averages around 400 Å^2 but can range from 200 to 1,200 Å^2. These areas are not correlated to the sizes of the proteins and only weakly to their apolar surface fraction. Ala, Lys, and Pro have dominating contributions to the apolar surface for smaller patches, while those of the hydrophobic amino acids become more important as the patch size increases. The hydrophilic amino acids expose an approximately constant fraction of their apolar area independent of patch size; the hydrophobic residue types reach similar exposure only in the larger patches. Though the mobility of residues on the surface is generally higher, it decreases for hydrophilic residues with increasing patch size. Several characteristics of hydrophobic patches catalogued here should prove useful in the design and engineering of proteins.**

Key words: molecular recognition, molecular surface, lipophilicity

INTRODUCTION

The molecular surface of proteins is of prime interest in the study of their physical and structural characteristics as well as their biological role. There must be constraints on the level of hydrophobicity at the protein surface. This is amply demonstrated by integral membrane proteins that dissolve in lipid bilayers but unfold and aggregate in aqueous media.[1] Another example is crambin, a very hydrophobic protein that is insoluble.[2]

Such phenomena are manifestations of the hydrophobic effect: the lack of solubility of apolar compounds in water due to the strong cohesion forces in the aqueous medium that arise from strong hydrogen bonding. Water molecules in contact with an apolar surface will either have to sacrifice hydrogen bonds or, by maintaining them, will lose entropy since there are fewer configurations that allow the maximal number of hydrogen bonds.[3]

The hydrophobic effect is thought to be the main determinant of protein folding.[4,5] The protein achieves in the folding process a minimum in solvent-exposed hydrophobic area and simultaneously an optimal solvent accessibility of polar atoms.[6] Burying hydrophobic groups in the protein core shields them from water. Apolar groups can also avoid interaction with the solvent through intermolecular association such as in subunit oligomerization.[7–9] Such interactions may not be intentional; in sickle cell hemoglobin a mutation of an exposed glutamic acid into a valine results in polymerized fibers that disable normal erythrocyte function.[10]

Hydrophobic contacts are also exploited in more transient molecular associations such as in the recognition and binding of substrates by enzymes. They are also found in many systems involving electron transfer proteins. For example, plastocyanin interacts with chlorophyll P700 of photosystem I through a hydrophobic patch[11]; azurin, with both nitrite reductase and cytochrome c_{551}[12]; cytochrome c, with cytochrome c oxidase[13]; and amicyanin, with methylamine dehydrogenase.[14] Kinases recognize and interact with their target protein partly via hydrophobic contacts.[15] Calmodulin encloses the apolar portion of its target peptide with two large hydrophobic patches on each of its two domains.[16]

Although the relevance of surface hydrophobic patches to the stability and function of proteins is evident, they have received little systematic study with the exception of subunit interfaces.[7–9,17,18] Though these regions are more hydrophobic than the rest of the protein surface, they are not representative of the surface of functional proteins since they are not normally exposed to the solvent.

Recently, Covell and co-workers[19] have described a method of finding regions on protein surfaces where a residue-based hydrophobicity potential is high. These regions generally coincided with binding sites of ligands or other proteins. This conclusion is consistent with ours (preceding paper) and supports the view that hydrophobicity is a key determinant in the recognition of cognate molecules.

In the preceding article we have described an algorithm for identifying hydrophobic patches. The

Received July 18, 1995; revision accepted January 22, 1996.
Address reprint requests to Patrick Argos, European Molecular Biology Laboratory, Meyerhofstrasse 1, Postfach 10.2209, 69012 Heidelberg, Germany.

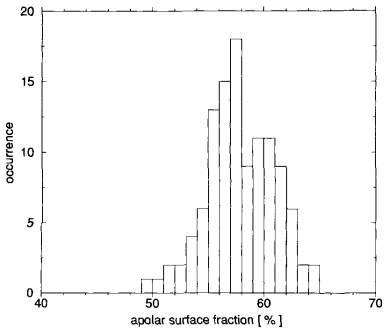

Fig. 1. Distribution of protein surface hydrophobicities in 112 monomeric proteins, expressed as the percentage of apolar surface, relative to the total solvent-accessible surface of each protein. Bins are 1% wide.

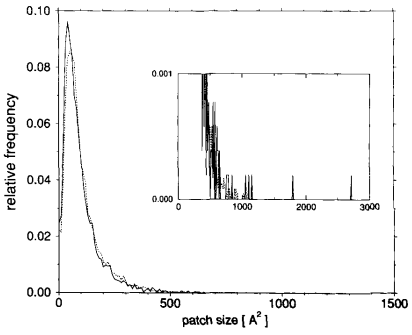

Fig. 2. Distribution of patch sizes in 112 proteins. The solid line represents the total number of observed patches in a bin of width 10 Å^2; the dashed line depicts the distribution found on the randomized protein surfaces. **Inset:** Details for the tail of the distribution.

method delineates contiguous solvent-accessible surfaces composed solely of carbon and sulphur atoms and delimited by polar atoms separated by less than two solvent diameters. Using this definition, we present here an overview of general trends concerning hydrophobic patches as found in a study of 112 unrelated monomeric proteins. Questions relating to size, composition, sequence distance, and secondary structure as well as correlations with crystallographic temperature factors will be addressed. The resulting patch characteristics should prove useful in protein engineering and design.

MATERIALS AND METHODS

Selection of Proteins

A maximal subset of protein tertiary structures was constructed by the method of Heringa et al.[20] The structures were characterized by a resolution better than 2.5 Å, sequence identities less than 40% in all pairwise alignments, chain length ≥ 45, and no missing coordinates for side chain atoms. The similarity threshold guarantees that more than one-half of all residues have been substituted. The surface residues of importance in this work are expected to be even more mutable. The majority of the protein structures had an R-factor better than 20%. Oligomeric proteins were excluded from the set. The protein structures were obtained from the Protein Data Bank[21] (PDB). The entry codes of the PDB files used are 155C, 1AAJ, 1AAK, 2ABK, 1ACX, 1ALB, 1ALC, 1ALD, 1ARB, 1ARP, 1ATR, 2AYH, 1BBC, 1BLC, 1BTC, 1CAA, 1CAJ, 1CDG, 1CLL, 1CMS, 1CPT, 1CRL, 1CTF, 1CTY, 2CY3, 1DHR, 1DOG, 1DRF, 2DRI, 1ECA, 1EDB, 2END, 1EST, 1FAS, 1FD2, 1FDX, 1FKB, 1FNC, 1FX1, 1FXD, 1GAL, 1GBT, 1GKY, 1GOF, 1GPR, 1HBQ, 1HFI, 1HMY, 1HOE, 1HUW, 1HYP, 1ICM, 1LAA, 1LE4, 1LEC, 1MBA, 1MBC, 1MDC, 1NAR, 1OFV, 1OMD, 1ONC, 1PAL, 1PAZ, 1PE6, 1PII, 1PNC, 1POH, 1RBS, 1RHD, 1SGC, 1SGT, 1STY, 1TFD, 1TML, 221P, 2AAA, 2ACT, 2APR, 2BAT, 2CDV, 2CP4, 2CPL, 2FCR, 2FGF, 2FXB, 2HPR, 2LHB, 2LIV, 2MCM, 2MHR, 2PF2, 2PIA, 2PK4, 2REN, 2SAS, 31BI, 351C, 3ADK, 3B5C, 3CHY, 3CPA, 3DFR, 3FXN, 3GBP, 3LZM, 3PGK, 3TGL, 5ACN, 6NN9, 6RXN, and 7ABP.

Solvent, ions, and hydrogen atoms were excluded in the calculations while cofactors were included. For a number of oligomeric proteins, patches on the complete functional complexes were also analyzed; however, the results were equivalent to those obtained for the monomers and will not be described here.

Identification of Patches

The technique used to detect hydrophobic patches has been described in detail in the preceding paper; only a summary will be given here. We define a patch as a contiguous portion of solvent-accessible surface,[22] consisting solely of neighboring carbon and sulphur atoms and mostly bordered by nitrogen and oxygen atoms. The solvent-accessible surface is obtained by adding the radius of a solvent molecule ("probe") to the atomic radii and then calculating a dot surface representation with the fast method of Eisenhaber et al.[23] Atomic radii utilized were those of Eisenberg and McLachlan[6]; the solvent radius was taken as 1.4 Å. Preliminary patches are subsequently delineated by expanding the oxygen and nitrogen atoms by yet an additional 1.4 Å and tracing the contiguous apolar surface segments that result. The expansion of the polar atoms closes any hydrophobic surface passages that may connect large clusters of exposed apolar atoms, thereby allowing easy

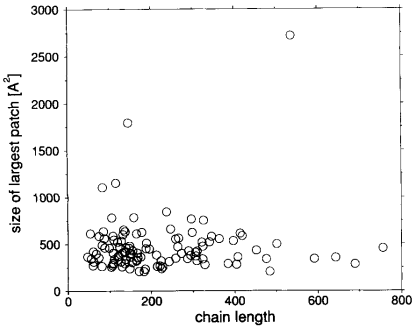

Fig. 3. Size of the largest patch per protein as a function of the chain length of the protein.

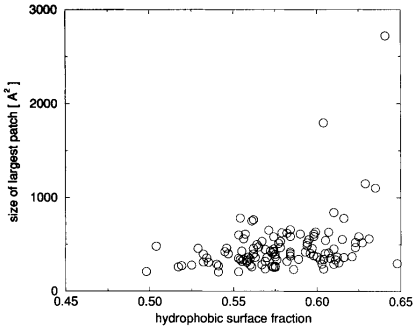

Fig. 4. Size of the largest patch per protein as a function of the apolar surface fraction of each protein.

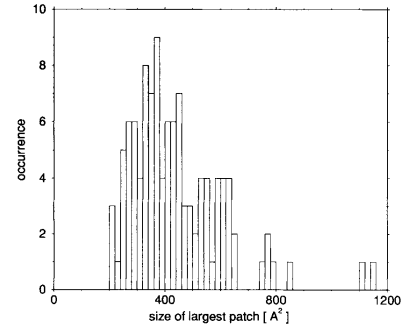

Fig. 5. Distribution of the area of the largest patch per protein. Bin size 10 Å². Two outliers (calmodulin, 1,794 Å² and lipase, 2,718 Å²) have been omitted.

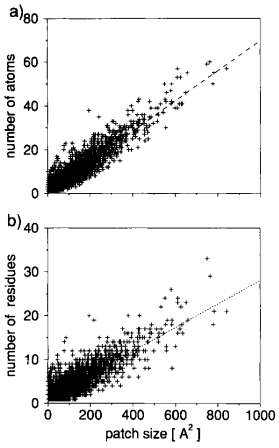

Fig. 6. Number of components per patch as a function of patch size. **a:** Number of atoms per patch. **b:** Number of residues per patch. Linear regression lines are drawn for both graphs.

recognition of the principal hydrophobic patches. To mitigate the loss of hydrophobic surface due to the expansion of polar atoms, each patch was then refined by adding to it any solvent-accessible apolar atoms directly adjacent to the preliminary patch atoms.

Randomization of protein surfaces for control purposes was performed as follows. All atoms were changed into carbon; then, atoms with solvent-accessible surface more than 10 Å² were randomly selected and converted into nitrogen and oxygen until the total of the original accessible surface area of such atoms was attained.

Unless stated otherwise, the results emanating from an analysis of patch characteristics were pooled into successive bins of 100 Å² for patch sizes up to 700 Å². The larger and fewer patches with surface area of 700 Å² and greater were put in a single bin. All areas given are for the solvent-accessible surface.

RESULTS AND DISCUSSION

Surface Hydrophobicity

The distribution of hydrophobic protein surface for the structures in our set, defined as the percentage of solvent-accessible hydrophobic surface relative to the protein's total solvent-accessible area, is shown in Figure 1. The apolar surface fraction lies between 49.8% and 64.8%, with an average of 57.8 ± 3%. For comparison, the hydrophobic surface area of typical amino acids in an extended Ala-X-Ala tripeptide are 32.5% (Gln), 55.8% (Thr), and 67.7% (Tyr). This demonstrates that a significant fraction of protein surfaces is surprisingly apolar.

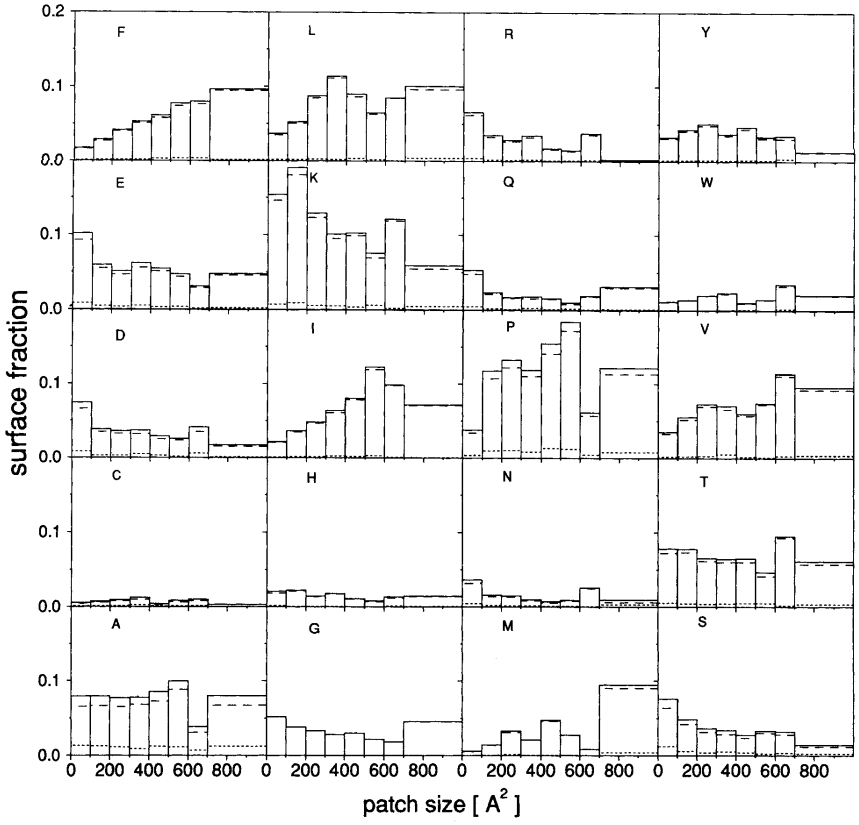

Fig. 7. Fractional apolar surface area contributions of amino acids as a function of patch size. Solid lines, total contribution; dotted lines, main chain contribution; dashed lines, side chain contribution.

Size Distribution

In our sample of 112 proteins, a total of 5,173 hydrophobic patches was found. The distribution of their sizes is shown in Figure 2. It is evident that large patches are rarer than smaller ones. Although solvent exposure of hydrophobic surface is unfavorable, the reason for the decrease in patch size would appear to be statistical rather than energetic since the randomized distribution strongly resembles the actual one (Fig. 2). Large patches occur when a large surface region is devoid of exposed polar atoms. The likelihood of a region containing no polar atoms decreases with its area, and large patches should therefore occur less frequently than smaller ones.

Patches smaller than 40 Å^2 are less abundant, which is the likely result of two factors. The first is due to artefacts in the detection procedure. Very small portions of hydrophobic surface will not survive the polar expansion and will therefore never be assigned to patches. In addition, the larger preliminary patches are given precedence during the refinement stage, causing a bias toward larger patches. The other factor lies in the covalent structure of hydrophilic side chains, which often expose their carbon atoms together with their polar moieties. Nonetheless, the most frequent apolar patch size is 40 Å^2, which corresponds to the area of two methylene groups and which can be in contact with about 4

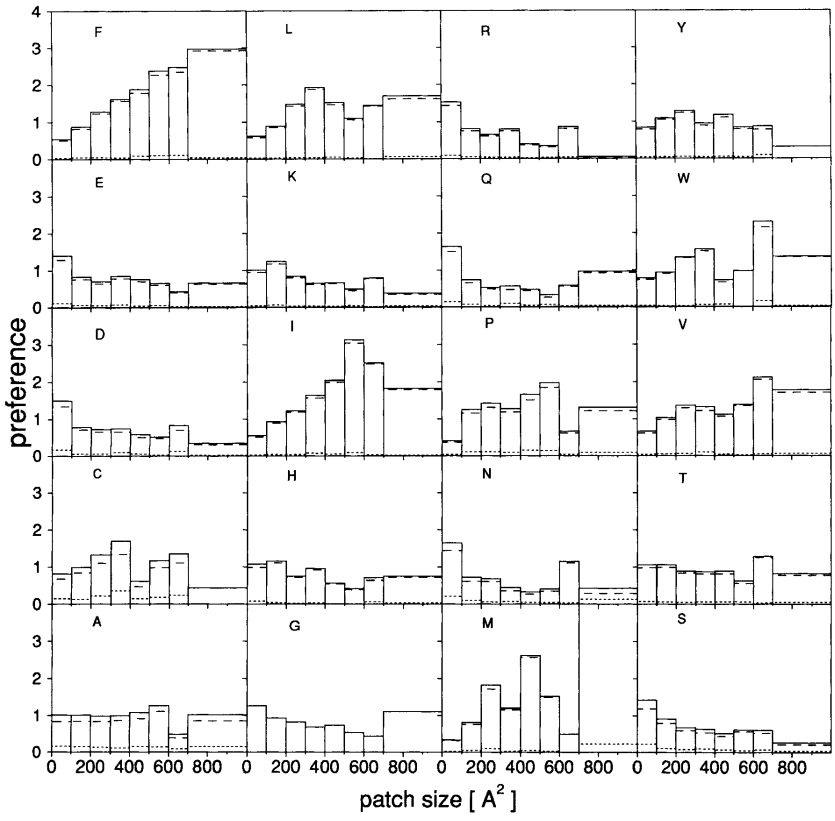

Fig. 8. Amino acid preferences for occurrence in patches of a given size, expressed as apolar surface fraction normalized by the apolar surface fraction of the corresponding amino acid over the entire protein surface. Subdivision over main chain and side chain atom contributions are as in Figure 7.

water molecules. Detection of such patches demonstrates the sensitivity of the method as well as the per-atom nature of usual surface patches.

Though large patches are less frequent, they are not absent. The size of the largest patch in each protein in relation to the protein size is depicted in Figure 3. No correlation is observed between the size of the protein and that of the largest patch. These results show that the fractional area of the largest patch relative to the total protein surface area decreases with protein size. This is quite unexpected, as the likelihood of the occurrence of a larger patch should increase with protein size. A larger patch should also in principle be tolerated more easily by a larger protein, which has more possibilities of compensating the energetic cost of exposing a large hydrophobic area. The size of the largest patch correlates only very weakly with the fraction of the protein's surface that is hydrophobic (Fig 4).

An explanation for the avoidance of particularly large patches is that they may associate if their sizes exceed a threshold that is independent of protein size. Such intermolecular association, which is a consequence of the hydrophobic effect, leads to aggregation, often detrimental to protein function. The same mechanism is utilized by proteins when the association is intended as in oligomerization.[7,8,24] Another possible explanation resides in local unfold-

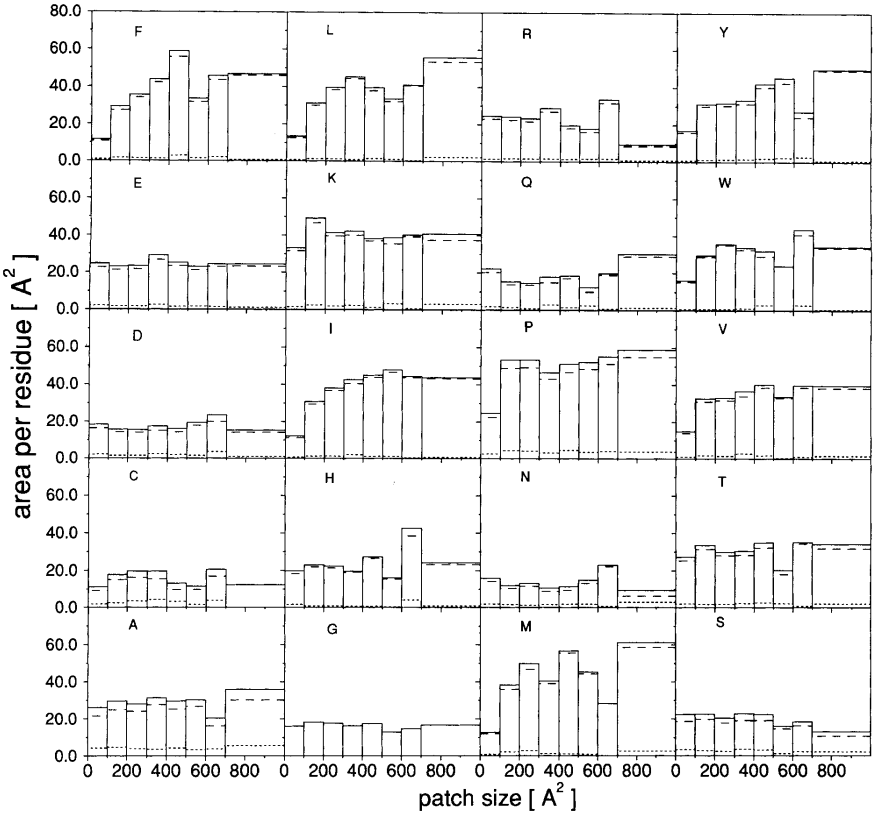

Fig. 9. Exposed hydrophobic area per single residue. Subdivision over main chain and side chain atoms are as in Figure 7.

ing processes that may ensue if patches exceed the upper limit; however, it is not possible presently to estimate a threshold above which such unfolding is triggered.

Figure 5 shows the distribution of the area of the largest patch per protein. The largest patches have surface areas roughly between 100 Å^2 and 600 Å^2 (95% of the sample, discarding the outliers lipase, 2,718 Å^2, and calmodulin, 1,794 Å^2). Such a distribution would suggest that a patch larger than 600 Å^2 may risk protein aggregation. The average largest patch measures 473 ± 300 Å^2, but the peak of the distribution occurs at 380 Å^2 due to its asymmetry. With a hydrophobic solvation free energy of 16 cal/mol/Å^2, the energetic cost of exposing a patch of 380 Å^2 to the solvent is a considerable 6.1 kcal/mol.[6,25] Often a specific function can be attached to patches of this size (preceding paper).

Figure 6 plots the numbers of residues and atoms involved in hydrophobic patches as a function of their size. The relationship is clearly linear, indicating 0.069 atom/Å^2 (14.6 Å^2/atom) and 0.027 residue/Å^2 of patch area (37.2 Å^2/residue) respectively. The latter value varies little across amino acid types and patch sizes (not shown). Although a patch certainly will increase in size if some of its atoms become more solvent accessible, this effect is not noticeable due to averaging effects. As will be seen below, the average exposure per residue is roughly constant for the hydrophilic amino acids but shows an increase with patch size for the hydrophobic residue types.

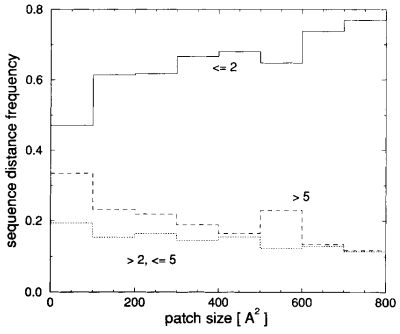

Fig. 10. Distribution of sequential distances among residues in one patch as a function of patch size (see text for an explanation).

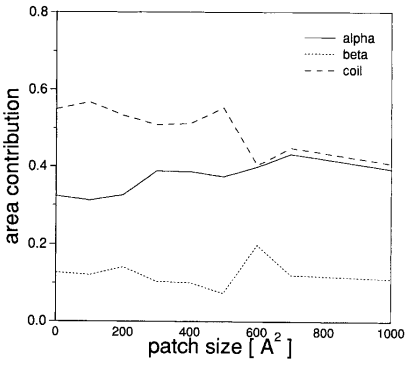

Fig. 11. Apolar surface area fractions contributing to hydrophobic patches by residues in different secondary structural states.

Composition

The surface area contribution of each amino acid type to hydrophobic patches in relation to patch size is depicted in Figure 7. The figure also indicates the partition over side chain and main chain atoms. Ala is a significant contributor, especially when compared with the large hydrophobic amino acids; Gly is also relatively important. The idea that patches should be composed predominantly of hydrophobic residues is therefore untenable. Nevertheless, proteins do make more use of hydrophobic residues (Leu, Ile, Phe, Val, Met) to assemble progressively larger patches, while contributions of hydrophilic amino acids decrease accordingly. The main chain contributions are low and roughly constant. The large contributions of Lys and Pro reflect their high average solvent accessibility, which can be rationalized. Lysine usually exposes much of its carbon-rich side chain, presumably to retain the entropy associated with side chain mobility. Proline is mostly found at turns of the polypeptide chain, where it is of necessity more exposed.

The preferences of amino acid types to occur in patches of a particular size can be defined as the apolar surface fraction of the amino acid in the size class normalized by the surface fraction of the amino acid over the entire protein surface, thus correcting for the size and abundancy of the amino acid. A figure larger than unity implies preference; one that is lower implies avoidance. As can be seen in Figure 8, the larger hydrophobic residues favor the larger patches, whereas the smaller and hydrophilic residues prefer smaller patches. Although this is expected given the large apolar surface area of the former, resulting from many bonded carbon atoms, it may be somewhat artefactual: the presence of a polar residue in a large patch implies the presence of a polar atom that could serve to split a large patch into smaller ones. This effect must be small, however, since Tyr and Trp have polar side chain atoms and nonetheless conform to the trend. It is clear that Phe, Leu, Ile, and Met are the preferred hydrophobic residues, while Pro and Tyr head the polar list.

Figure 9 shows the hydrophobic exposure per single residue for each amino acid type. The smaller and hydrophilic amino acids display an apolar surface area roughly constant with patch size. In contrast, the hydrophobic amino acids show distinct increases up to patch sizes of around 500 Å^2, beyond which they remain relatively constant. The exposure increases to around 40% of the area in the extended conformation, as is the case for hydrophilic residues. Apparently, a maximum has been attained at this patch size; for yet larger patches, the growth in patch area must come from the addition of more residues to the patch rather than from more area per residue. Hydrophilic residues including His are already maximally exposed in the smaller patches.

Sequential Distance

Hydrophobic patches could be composed predominantly of residues near the primary structure or, conversely, involve the clustering of groups distant along the sequence. For a patch comprised of n residues ordered according to their sequence position number, there are n—1 sequence distances, defined as the difference in sequence position number between successive residues. The distances are classified as short (1 or 2), medium (3–5), or long range (> 5). About 60% of the possible distances are short range (Fig. 10), suggesting that patches consist of several sequentially distant stretches of successive residues, brought together in space. For example, if 80% of the distances in a 20-residue patch are short range, then this could be the result of the coales-

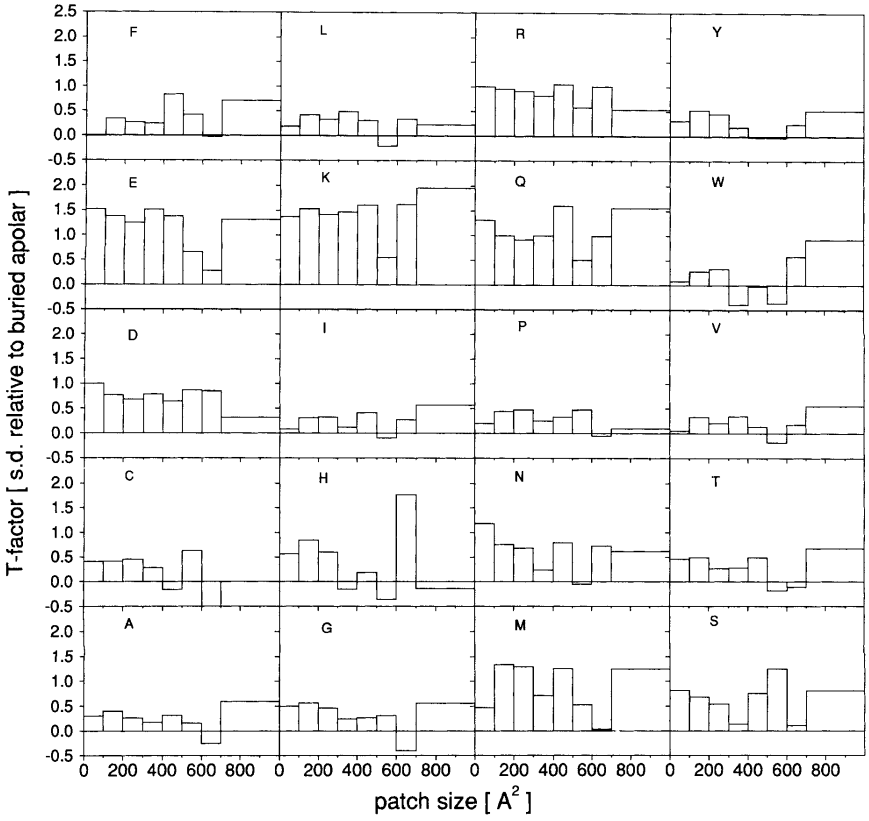

Fig. 12. Mobilities of atoms contributing to patches, expressed in standard deviations above or below the structure's average crystallographic temperature factor of buried apolar atoms.

cence of four stretches of about five residues each. The chance that the apolar surfaces of sequentially close residues will be adjacent and thus form or enlarge a patch is high due to the correlation between sequential and spatial distance; this explains the prevalance of short-range separations. The short-range contribution increases with patch size, implying that addition of residues to a stretch already forming a patch is the dominant mode of obtaining larger patches.

Secondary Structure

The relative area contributed by residues in a particular secondary structural state (α-helix, β-sheet, and coil) is depicted in Figure 11. The graphs that reflect the surface composition in terms of secondary structure are largely featureless. Random coil, usually found at turns where it can expose hydrophobic surface, prevails. Correspondingly, β structure, often buried in the protein core, contributes least to the accessible surface. Although active sites often involve connecting loops having coil structure[26] and also frequently display large hydrophobic patches, this correlation is not present in the patch size dependence on the area contributions over the three secondary structural states.

Mobility

The average mobility of atoms is mirrored in the crystallographic temperature factors that attempt to quantify the oscillation of atoms about their central location. They are not numerically comparable

across different protein structures and were therefore expressed in standard deviations above or below the average temperature factor of apolar atoms of residues exposed less than 10 Å^2. Since side chains at the protein surface are generally more mobile than those in the hydrophobic core, the atomic movements in hydrophobic patches should deviate positively from this average, as is indeed observed in Figure 12. The mobility of hydrophobic amino acids (Leu, Ile, Val, Phe) is lower than that of the hydrophilic ones. This may be due to the dynamic nature of solvation, which is stronger for polar atoms. For many of the polar (Thr, Tyr, Asn), charged (Arg, Glu), and generally hydrophobic residues containing a polar atom (Trp, Tyr), the mobilities decrease as the patch size grows. Their mobility may be restricted increasingly by the presence of the less mobile hydrophobic side chains, which contribute more to the larger patches (Fig. 7). Though the mobility of patch residues is clearly greater than that of buried atoms, it is also less than that of polar atoms (data not shown).

CONCLUSIONS

The apolar surface fraction of proteins lies around 58%, with extrema at 50 and 65%. Large patches occur less frequently than smaller ones. The sizes of the largest patch in each individual protein range from 200 to 1,200 Å^2, the commonest having an area around 400 Å^2. These areas are not correlated to the sizes of the proteins and only marginally to their apolar surface fraction. The observed upper limit on hydrophobic patch size should prove significant in the design and engineering of proteins. The area contributions to patches are dominated by Ala, Lys, and Pro, but those of the hydrophobic amino acids are also significant and increase with patch size. The hydrophilic amino acids expose approximately 40% of their apolar area relative to their extended state, independent of patch size. The hydrophobic amino acids reach a similar value, but only for patches of 500 Å^2 and larger. Residues in patches are mostly neighboring in sequence. The mobility of residues on the surface is higher than that of buried groups but shows a decrease with increasing patch size for hydrophilic residues, interpreted as a hindering by the less mobile hydrophobic side chains.

ACKNOWLEDGMENTS

We thank Jaap Heringa and Frank Eisenhaber for useful suggestions and a critical reading of the manuscript and André Juffer for stimulating discussions. The authors also express their gratitude to the crystallographers who deposited their data in the Protein Data Bank. Finally, we are deeply indebted to Larry Wall, Richard Stallman, Paul Turner, Roger Sayle, and countless others for making freely available their superb software tools used in this work.

REFERENCES

1. Van Renswoude, J., Kempf, C. Purification of integral membrane proteins. Methods Enzymol. 104:329–334, 1984.
2. Teeter, M.M., Hendrickson, W.A. Highly ordered crystals of the plant seed protein crambin. J. Mol. Biol. 127:219–233, 1979.
3. Tanford, C. The Hydrophobic Effect. New York: Wiley, 1980.
4. Kauzmann, W. Some factors in the interpretation of protein denaturation. Adv. Protein Chem. 14:1–63, 1959.
5. Dill, K.A. Dominant forces in protein folding. Biochemistry 29:7133–7155, 1990.
6. Eisenberg, D., McLachlan, A.D. Solvation energy in protein folding and binding. Nature, 319:199–203, 1986.
7. Chothia, C., Janin, J. Principles of protein–protein recognition. Nature 256:705–708, 1975.
8. Argos, P. An investigation of protein subunit and domain interfaces. Prot. Eng. 2:101–113, 1988.
9. Janin, J., Miller, S., Chothia, C. Surface, subunit interfaces and interior of oligomeric proteins. J. Mol. Biol. 204: 155–164, 1988.
10. Fermi, G., Perutz, M.F. Atlas of Molecular Structures in Biology 2. New York: Clarendon Press, 1981.
11. Guss, J.M., Freeman, H.C. Structure of oxidized poplar plastoycynanin at 1.6 Å resolution. J. Mol. Biol. 169:521, 1983.
12. van de Kamp, M., Silvestrini, M.C., Brunori, M., van Beumen, J., Hali, F.C., Canters, G.W. Involvement of the hydrophobic patch of azurin in the electron-transfer reactions with cytochrome c551 and nitrite reductase. Eur. J. Biochem. 194:109–118, 1990.
13. Pelletier, H., Kraut, J. Crystal structure of a complex between electron transfer partners, cytochrome c peroxidase and cytochrome c. Science 258:1748–1755, 1992.
14. Chen, L., Durley, R.C.E., Mathews, F.S., Davidson, V.L. Structure of an electron transfer complex: Methylamine dehydrogenase, amicyanin and cytochrome c551. Science 264:86–89, 1994.
15. Taylor, S.S., Knighton, D.R., Zheng, J., Ten Eyck, L.F., Sowadski, J.M. Structural frame work for the protein kinase family. Annu. Rev. Cell Biol. 8:429–462, 1992.
16. Harpaz, Y., Gerstein, M., Chothia, C. Volume changes on protein folding. Structure 2:641–649, 1994.
17. Hubbard, S.J., Argos, P. Evidence on close packing and cavities in proteins. Curr. Opin. Biotechnol. 6:375–381, 1995.
18. Korn, A.P., Burnett, R.M. Distribution and complementarity of hydropathy in multisubunit proteins. Proteins 9:37–55, 1991.
19. Young, L., Jernigan, R.L., Covell, D.G. A role for surface hydrophobicity in protein–protein recognition. Prot. Sci. 3:717–729, 1994.
20. Heringa, J., Sommerfeldt, H., Higgins, D., Argos, P. OBSTRUCT: A program to obtain largest cliques from a protein sequence set according to structural resolution and sequence similarity. CABIOS 8:599–600, 1992.
21. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F.J., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. A computer based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
22. Lee, B., Richards, F.M. The interpretation of protein structure: Estimation of static accessibility. J. Mol. Biol. 119: 537–555, 1971.
23. Eisenhaber, F., Lijnzaad, P., Argos, P., Sander, C., Scharf, M. The double cubic lattice method: Efficient approaches to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies. J. Comp. Chem. 16:273–284, 1995.
24. Janin, J., Chothia, C. The structure of protein–protein recognition sites. J. Biol. Chem. 26:16027–16030, 1990.
25. Juffer, A.E., Eisenhaber, F., Hubbard, S.J., Walther, D., Argos, P. Comparison of atomic solvation parametric sets: Applicability and limitations in protein folding and binding. Prot. Sci. 4:2499–2509, 1995.
26. Brändén, C.-I. Relation between structure and function of alpha/beta proteins. Q. Rev. Biophys. 13:317–338, 1980.

5

INTERFACE PATCHES

P. Lijnzaad and P. Argos.
Hydrophobic patches on protein subunit interfaces: characteristics and prediction.
Proteins **28**:333-343 (1997)

Hydrophobic Patches on Protein Subunit Interfaces: Characteristics and Prediction

Philip Lijnzaad[1] and Patrick Argos[2]*
[1]*EMBL Outstation Hinxton (EBI), Wellcome Trust Genome Campus, Hinxton, United Kingdom*
[2]*European Molecular Biology Laboratory, Heidelberg, Germany*

ABSTRACT **Hydrophobic patches, defined as clusters of neighboring apolar atoms deemed accessible on a given protein surface, have been investigated on protein subunit interfaces. The data were taken from known tertiary structures of multimeric protein complexes. Amino acid composition and preference, patch size distribution, and patch contact complementarity across associating subunits were examined and compared with hydrophobic patches found on the solvent-accessible surface of the multimeric complexes. The largest or second largest patch on the accessible surface of the entire subunit was involved in multimeric interfaces in 90% of the cases. These results should prove useful for subunit design and engineering as well as for prediction of subunit interface regions. Proteins 28:333–343, 1997.**

Key words: protein structure; oligomeric structure; subunit interface; molecular recognition

INTRODUCTION

The biologically active form of many protein molecules is a complex of two or more polypeptide chains. The functions of such multimeric subunit associations include allosteric control mechanisms, signal transduction, binding of symmetrical substrates, interaction with active sites, and reduced diffusional mobility of the complex.

Molecular association involves loss of entropy for the monomers or subunits. In a dimeric complex, the negative entropy change arises from the loss of six motional degrees of freedom with additional losses in interface side-chain mobilities. Janin and coworkers[1] have estimated the entropic cost at 20–30 kcal/mol and have suggested that the burial, upon complexation, of exposed hydrophobic surface area (the hydrophobic effect) is the main driving force for oligomerization. The interface region of monomers has also been found to be more hydrophobic than their solvent-exposed surface, which supports this conclusion.[2] Principal studies of subunit association surfaces have included those of Argos,[2] Janin et al.,[1] Korn and Burnett,[3] and Jones and Thornton.[4]

We have recently described a method for detecting explicit, contiguous patches of hydrophobic surface and have applied it to a set of monomeric proteins. The protein's surface can be viewed as several clusters of neighboring carbon and sulfur atoms (hydrophobic surface) surrounded by polar or charged nitrogen and oxygen atoms (hydrophilic surface). The large hydrophobic patches are often connected by narrow hydrophobic "channels" resulting from the extended contiguity of touching apolar atoms. These channels can be eliminated to elicit the significant hydrophobic patches.[5] In the current investigation, we examine the hydrophobic patches found on the surface of subunit interfaces and address questions concerning the size and shape of the patches, their coincidence with the interface and contact with patches on the partner subunit, their amino acid compositional biases, and the prediction of their involvement in oligomeric association. This insight into the details of hydrophobic subunit interaction should prove valuable in protein engineering, design, and docking.

METHODS

A set of multimeric proteins was obtained from the Brookhaven protein tertiary structure data bank[6,7] by selecting entries, with the aid of the SRS information search system,[8,9] that contain more than one polypeptide chain, each longer than 50 residues, or display in the REMARK-section of the entry file words indicative of an oligomeric protein such as dimer, trimer, or tetramer. From this initial set, a subset containing the maximal number of protein chains with all pairwise sequence alignments of $<45\%$ in residue identity was constructed by the method of Heringa and coworkers.[10] For the latter set, we selected those structures where more than 1000 Å^2 of a surface is buried per chain upon complexation. This yielded 59 protein complexes of which 41 were dimeric, 2 trimeric, 13 tetrameric, 2 pentameric, and 1 hexameric, totaling 156 polypeptide chains. The PDB entry codes and chain identifiers (listed successively as single letters after the

*Correspondence to: Dr. Patrick Argos, European Molecular Biology Laboratory, Meyerhofstrasse 1, Postfach 10.2209, 69012, Heidelberg, Germany.
Received 11 June 1996; Accepted 27 February 1997

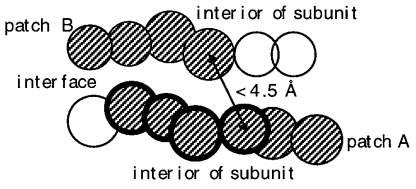

Fig. 1. The definition of patch overlap. Patches on two different subunit interfaces *A* and *B* are shown. Gray circles indicate hydrophobic atoms while white circles refer to hydrophilic atoms. Atoms of patch *A* with thick circumference indicate those that overlap with patch *B*. The sum of the full solvent accessibility of the latter atoms in the uncomplexed state is taken as the measure of the extent of overlap.

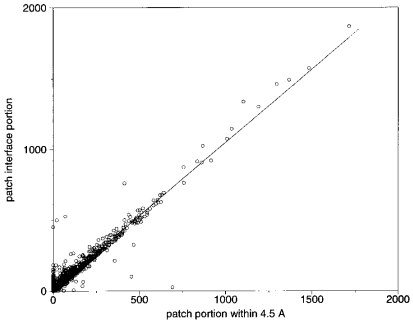

Fig. 2. A plot showing the actual loss of surface area upon subunit association versus the amount estimated on the basis of interatom contacts with distance criterion 4.5 Å. The linear regression line ($Y = 1.03 * X + 12.25$) has a correlation coefficient of 0.97 with the data.

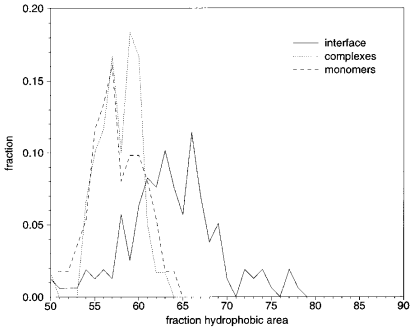

Fig. 3. A histogram of the percentage of solvent accessible surface that is hydrophobic. The solid line plot represents interface surface percentages, while the dotted and dashed lines apply to the hydrophobic fraction on the exterior surface of multimeric complexes studied here and to the monomeric proteins from our previous study.[5]

hyphen) of the proteins used in this work are as follows:

1ABR-AB, 1ALK-AB, 1AOR-AB, 1CAU-AB, 1CHM-AB, 1CMB-AB, 1EAP-AB, 1EBH-AB, 1FIA-AB, 1GER-AB, 1GLP-AB, 1GST-AB, 1HLD-AB, 1MRR-AB, 1NHK-LR, 1NSB-AB, 1POW-AB, 1PP2-LR, 1PVU-AB, 1PYD-AB, 1SES-AB, 1SRI-AB, 1TKB-AB, 1TPH-12, 1TPL-AB, 1VAA-AB, 2BBQ-AB, 2CST-AB, 2FB4-HL, 2FBJ-HL, 2NAD-AB, 2POL-AB, 2RSP-AB, 2SCP-AB, 2TMD-AB, 2UTG-AB, 3AAH-AC, 3SC2-AB, 4HVP-AB, 5RUB-AB, 8CAT-AB; 1BBP-ABCD, 1BOV-ABCDE, 1COM-ABC, 1CPC-ABKL, 1EPT-ABC, 1GD1-OPQR, 1HDC-ABCD, 1HJR-ABCD, 1HSA-ABDE, 1HUC-ABCD, 1NBA-ABCD, 1PYA-ABCDEF, 1RAI-ABCD, 1SAC-ABCDE, 2BBK-HLJM, 2HNT-LCEF, 2MTA-HLAC, 3PGA-1234.

The method used to delineate the hydrophobic patches has been previously described by us.[5] The solvent accessible surface of a folded polypeptide chain can be considered to consist of clusters of neighboring hydrophobic (carbon and sulfur) atoms surrounded by contiguous strings of hydrophilic (nitrogen and oxygen) atoms. However, it is very likely that the hydrophobic clusters will be connected by—sometimes narrow—"channels" consisting of strings of hydrophobic atoms. These channels can be closed through the expansion of the hydrophilic atoms (only if solvent-accessible) by a certain amount. With the remaining preliminary hydrophobic patches so identified, the expanded hydrophilic atoms are then reduced to their normal size and one layer of apolar atoms (if they were previously covered by the expansion) are added to each of the preliminary patches.

Subunit interface atoms were defined as those that displayed a difference in solvent accessibility between the fully complexed and monomeric states. In all cases, accessibility was determined by the numerical method of Eisenhaber and coworkers,[11] with reliance on a solvent probe radius of 1.4 Å. For complexes consisting of more than two chains, all pairwise interfaces involving one subunit were considered together; for example, in the trimer $\alpha\beta\gamma$, the α interface was the union of the $\alpha\beta$ and $\alpha\gamma$ interaction surfaces.

A hydrophobic patch was designated an interface patch if more than a certain fraction of it was buried upon subunit association. In our sample of 156 polypeptide chains from 59 multisubunit proteins, a total of 8252 patches were found, of which 2556 had some overlap with the interface, 1891 had more than 50% of their solvent accessible surface involved in the interface, 1447 more than 80% overlap, and 759

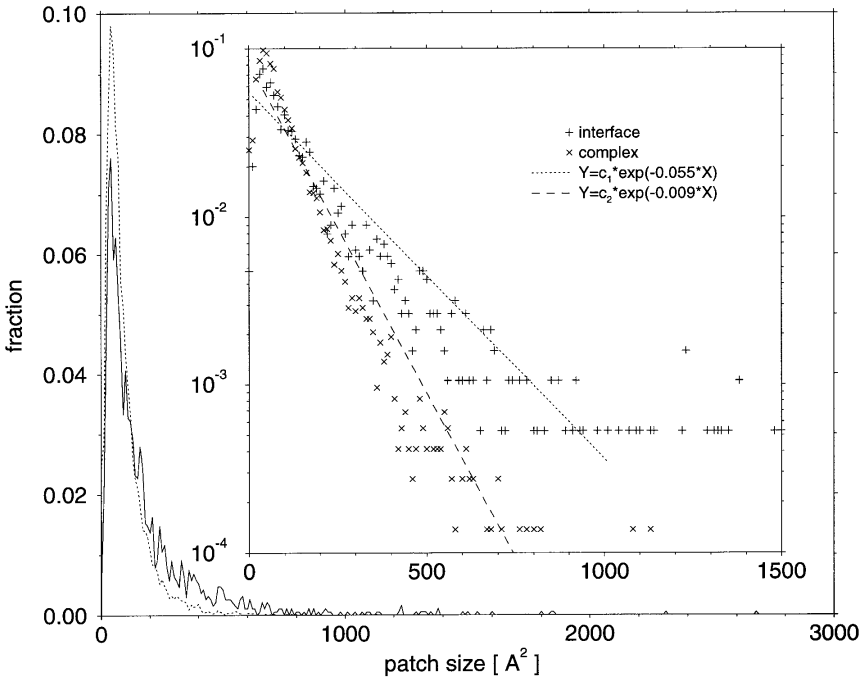

Fig. 4. The size distribution of hydrophobic patches with more than 50% of their individual accessible surface within the interface (solid line) and of hydrophobic patches on the exterior surface of multimeric protein complexes (dashed lines). The inset shows a logarithmic plot with regression lines given for both patch types (+ for interface and X for multimers). The correlation coefficient is 0.96 for the interface patches and 0.97 for patches on the exterior surface of complexes.

had a coincidence of more than 99%. In this work we defined an interface patch as one that has more than 50% of its area associated with subunit interface(s). When a different threshold is used, it will be stated explicitly.

To measure the overlap of interface patches from different subunits, we used the following criteria. For each atom of a patch A on one subunit, we consider it to display overlap with patch B on the other subunit if this atom is closer than 4.5 Å to any atom of patch B (Fig. 1). The contact distance 4.5 Å was employed as 4.0 Å represents about one carbon diameter with the extra 0.5 Å allowing for experimental error and hydrogen atoms. The sum of the (uncomplexed) atomic solvent-accessible surface areas of the atoms in patch A that, with the 4.5 Å criterion, are overlapped by atoms in patch B will be denoted overlap (A, B). Due to curvature effects, overlap (A, B) is generally not equal to overlap (B, A).

The sum of overlap (A, B_i) for patch A in overlap with several different patches B_i, divided by the entire area of patch A, is called the overlap factor $\Phi(A)$, that is,

$$\Phi(A) = \sum_i \frac{\text{overlap}\ (A, B_i)}{\text{area}\ (A)}$$

where $\Phi(A)$ is the fraction of a patch's area that overlaps with patches on the other subunit(s). Occasionally $\Phi(A)$ is larger than the area of patch A, as atoms of A can be counted doubly for different B_i. In this case the overlap factor was set to 1. Overall statistics were obtained by combining and averaging the results from overlap (A, B) and overlap (B, A).

Formally, the change in accessible surface on contact of every possible patch pair should be calculated; however, the computer calculations over all possible patch pairs and all interfaces were prohibitive, forcing the approximation through atom contacts and accessible surface summation. Since the actual area contributed by a patch to the interface

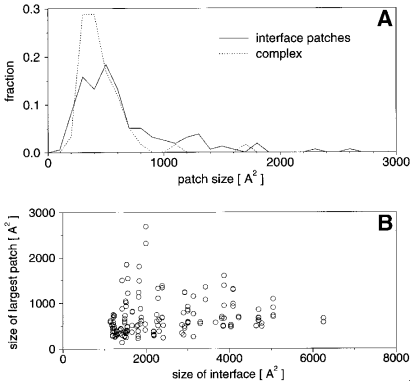

Fig. 5. **A:** The distribution of sizes of the largest hydrophobic interface patch in each subunit as well as for hydrophobic patches on the exterior surface of multimeric protein complexes. **B:** The size of the largest interface patch as a function of the size of the subunit interface. The latter is expressed as the total area loss upon subunit association per polypeptide chain.

had been determined through the monomeric/multimeric accessible surface calculations, the area of a given interface designated patch estimated by all contacting atoms from other subunit patches could be compared with the actual area. Figure 2 shows a plot of the two areas which display good correlation except for a few outliers.

In this work several distribution plots (histograms) are shown. In each case the vertical axis is labeled "fraction," while the horizontal axis is denoted by the name of a property examined over the protein monomers or multimeric complexes used in this work. The fraction label thus refers to the percentage (fraction) of all the proteins or patches studied that displayed particular values for a given structural characteristic labeled and plotted on the horizontal axis.

RESULTS AND DISCUSSION

Patch Size Distribution

Subunit interfaces are known to be more hydrophobic than the external protein surface.[2,3] This is confirmed in Figure 3, which shows histograms of the fraction of exposed (accessible) surface that is hydrophobic for several monomeric proteins (data from Ref. 12); for multimeric complexes used in this study; and for subunit interface surfaces. Regarding the fraction of the surface that is composed of hydrophobic atoms, the exterior surfaces of multisubunit complexes and monomeric proteins cannot be distinguished, while subunit interfaces show considerably elevated fractions. Monomeric and complex surfaces are all 50% to 65% hydrophobic, while interfaces mostly range between 58% and 70%.

The distribution of the sizes of hydrophobic patches is given in Figure 4 for interface and complex surfaces. Although the distributions generally appear similar, the plot for the exterior surface of complexes levels off sooner than that for interface patches, showing that the latter possess more large patches. This is expected, given the more hydrophobic subunit interfaces. Both distribution curves behave roughly exponentially, as demonstrated in the logarithmic plot of the inset of Figure 4. Regression lines fitted to the logarithm of the original distributions show good correlation. The fivefold difference in the factors of the fitted exponents confirm that there are more large hydrophobic patches on subunit interfaces.

Since there are more large patches on interfaces, it is possible that the largest patch in each interface is, on average, larger than the largest patch found on the exterior surface of the complex. This, however, is not the case, as shown in Figure 5a, which displays histograms of the size of the largest patches in each subunit or protein for interfaces and for complexes, respectively. Both histograms display similar peaks. There is also no relationship between the size of the interface and the size of the largest patch (Fig. 5b). This behavior resembles that found in monomeric proteins, where no correlation was found between the size of the protein and that of the largest patch.[12]

Number of Patches per Interface

In Figure 6 the number of hydrophobic patches per interface is shown according to various patch size thresholds. As expected, the average number of hydrophobic patches per interface decreases as the patch size considered increases. The number of patches larger than 100 Å^2 ranges between 2 and 14 for most interfaces, while, for patches larger than 300 Å^2, the most common number of patches per interface has reduced to just one. Patches larger than 600 Å^2 are found in only 40% of the interfaces.

Patch Overlap With Interface

Do the largest hydrophobic patches on subunit surfaces lie predominantly within the interface region, and do interface patches mostly coincide with the interface? These questions are addressed in Figure 7, which shows the distribution of overlap of patches with the interface, both for all patches and for the largest patch on the entire and isolated subunit surface of each chain considered. The large fraction of patches in the bin with 0–10% of their patch area overlapping the interface encompasses all the noninterface patches, many of which are

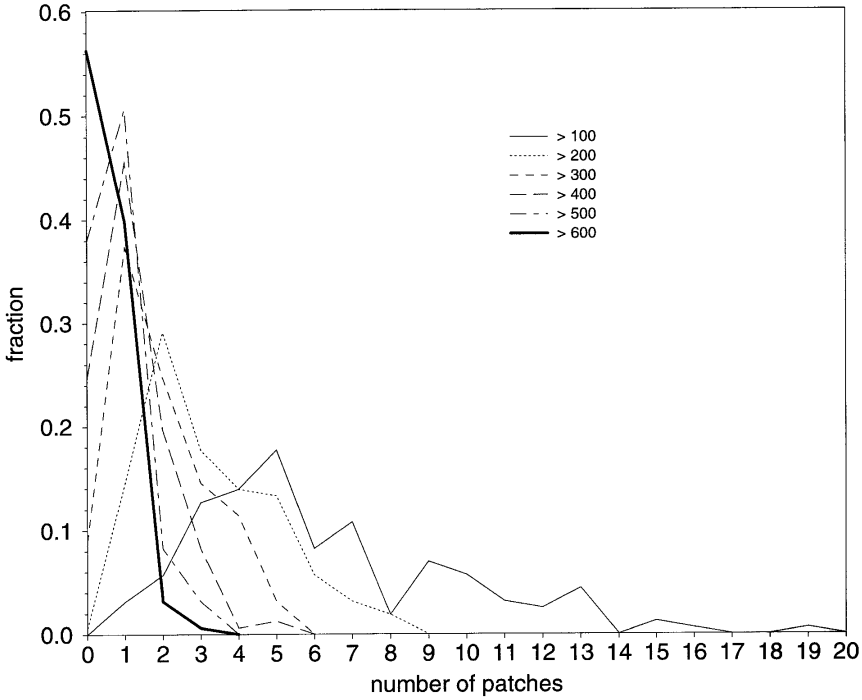

Fig. 6. A relative histogram of protein subunits having a certain number of hydrophobic interface patches as a function of the minimum size of the patches considered.

relatively small. The distribution peaks at the 90–100% bin, corresponding to nearly 40% of the proteins examined. Based on the current sample, this shows that if the largest patch has any overlap with the interface, it is most likely to have in fact an overlap of more than 90%. Also, the cumulative distribution shows that more than three-fourths of the largest patches on the entire subunit surface have an overlapping surface of more than 50% with the interface. Thus, for the largest patches, the transition from interface to exterior surface is rather abrupt, and this may well be an important factor in the process of recognition and complexation. Nonetheless, there is a reasonable fraction of largest patches not involved in the interface (16%), although they may have a functional role.[12]

Patch Rank and Occurrence in Interface

Large patches are frequently found in the interface, but is the largest patch of a subunit always involved in the interface? Figure 8 shows that this is the case for 75% of the multimers in our data set. From the cumulative histogram given in the inset, it can be seen that either the largest or the second largest hydrophobic patch of the entire subunit surfaces lies within the interface for 90% of the multimeric proteins examined. A patch is deemed to lie within an interface if more than 50% of its total surface is involved in subunit association. Young and coworkers[13] have noted a correlation between points of high hydrophobic potential and protein binding sites. Our finding should be very useful in the study of subunit-subunit docking, as it greatly limits the search space for recognition studies. Amino acid residues composing the largest patch are thus a likely target to alter oligomerization characteristics.

Patch Complementarity

When subunits associate, they mutually bury hydrophobic surface area. It should be useful to know

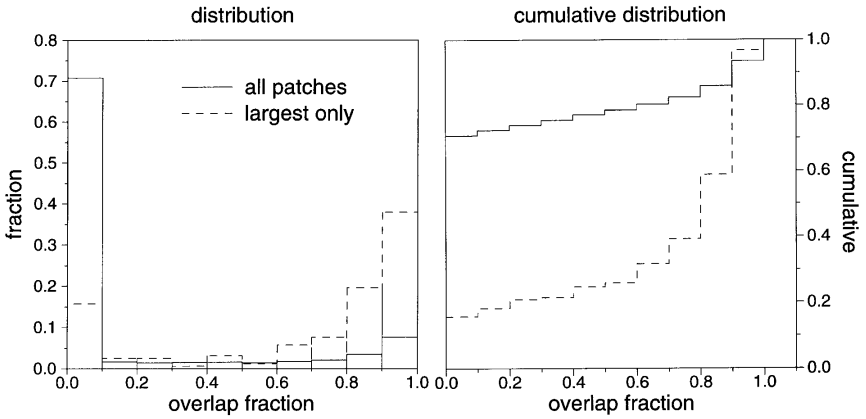

Fig. 7. The distribution of the overlap fraction of the total patch surface with the interface surface for all patches and for the largest patches only on the accessible surface of the associated subunits.

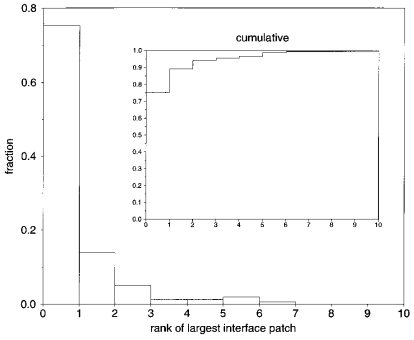

Fig. 8. The distribution of the rank number for the largest interface patch on each subunit. The cumulative distribution plot is also shown. The number taken for each subunit is the rank among all hydrophobic patches found on the entire subunit surface. The inset shows the cumulative distribution.

to what extent the larger hydrophobic patches on either subunit are brought into contact with each other upon association. Good "large patch complementarity" could be exploited in subunit–subunit docking studies. We have investigated patch complementarity using the measures described in the Methods section; for simplicity, only dimers were considered, since higher order multimers could well involve large patches being covered by those from several different subunits.

The distribution of the overlap factor $\Phi(A)$ (see the Methods section) is shown in Figure 9, which depicts the behavior of the overlap factor when selecting patches of various minimal sizes. For example, considering only patches 600 Å^2 or greater in surface area, over 60% of these A patches have only 0% to 10% of their total surface covered by the corresponding B patches of the same minimal size. The average fraction of all considered patches in Figure 9 are plotted according to the occurrence of their covering factor in the 10% ranges. For the largest patches (more than 600 Å^2 in area), only about 16% of them are covered to an extent of 60% or more of their surfaces by patches of similar size (see cumulative distribution, Fig. 9). When all patch sizes are used, the patch fraction is reasonably similar across the various overlap ranges as expected. As the minimum thresholds increase, the patch fraction without or with little contact increases. We therefore conclude that there is no significant preference for large patches on one subunit to meet and bury large patches on the other subunit.

Similar conclusions arise from a study of the number of patches that are in contact with those of any size on the other subunit. If the overlap $A(B)$ exceeded 25% of the (uncomplexed) accessible area of patch A, we counted it as a contact. The resulting distribution of the number of contacts per patch is shown in Figure 10. Most patches have just one contact; for patches larger than 400 Å^2, most do not have any overlap exceeding 25% of their area.

Shape of Patches

To assess the general form or shape of the hydrophobic interface patches, we fitted an ellipsoid to each patch by using the method of Taylor and

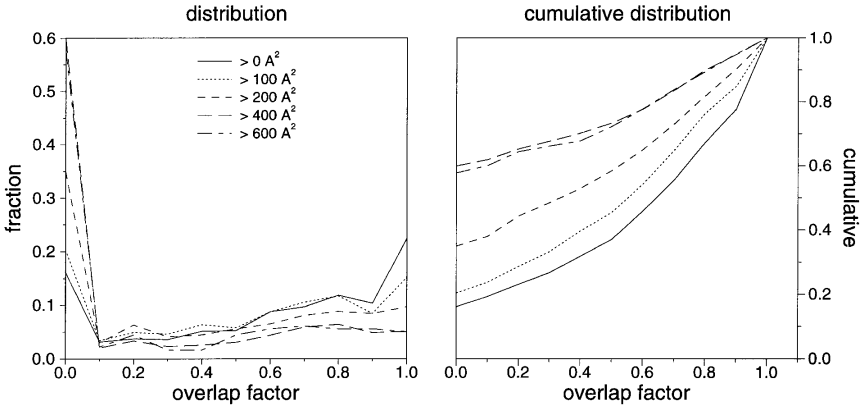

Fig. 9. Plot of the frequency versus the fraction of the interface patch surface on one subunit covered by interface patches on the other subunit, both patches being of an indicated minimum surface size. The cumulative distribution is also shown. The overlap fraction is plotted in ranges of 10%, such that, for example, at the overlap point 0.0 the fraction of cases is given for an overlap range 0 to <10%; at overlap fraction 0.1, the fraction of all cases is plotted for the 10 to <20%; and so forth to the last point plotted at 1.0 overlap fraction for 100% surface coverage.

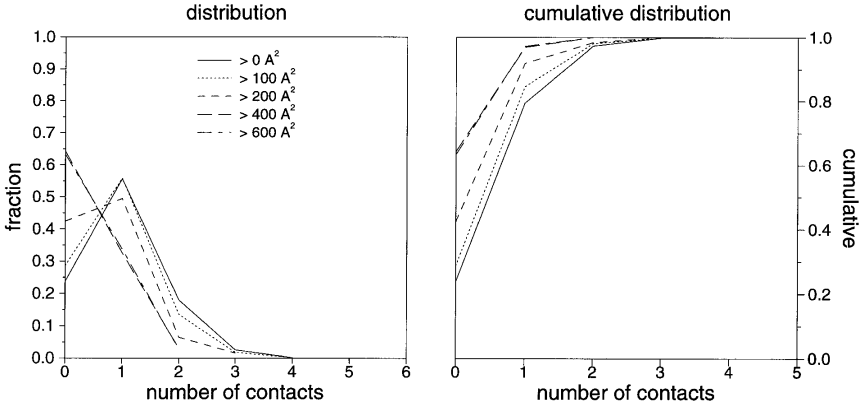

Fig. 10. The number of contacts across the interface for overlapping patches of varying sizes (see text for contact definition). The cumulative distribution is also given.

colleagues.[14] Their procedure yields the so-called equivalent ellipsoid, which follows the atom coordinate distribution more closely than does the inertial ellipsoid. The ellipsoidal axes so found were scaled by a common factor such that the centers of all the atoms that compose the patch were just contained in the ellipsoid. A scatter plot of the occurrence of the lengths of the two largest ellipsoid semiaxes is given in Figure 11. Most of the semiaxes have lengths below 15 Å, but some are much larger.

The ratio of the two largest semiaxes *a* and *b* quantifies the elongation of an ellipsoid; the inset of Figure 11 shows a histogram of the ratios. Close to 40% of the interface patches are roughly round in major cross section, having an *a*/*b* ratio of 1.5 or less; the rest can be described as elongated, with smaller

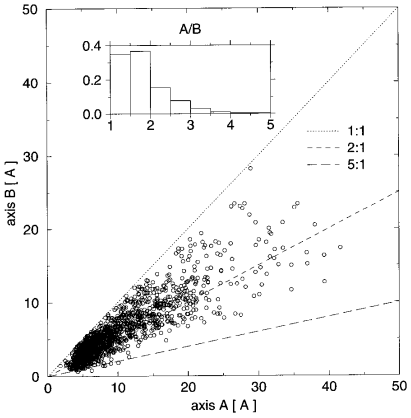

Fig. 11. A scatter plot of the occurrence of the two largest semiaxial lengths of equivalent ellipsoids fitted to the interface patches. The lines delineate the regions of the axial ratios noted in the plot. The inset shows the distribution of the ratios.

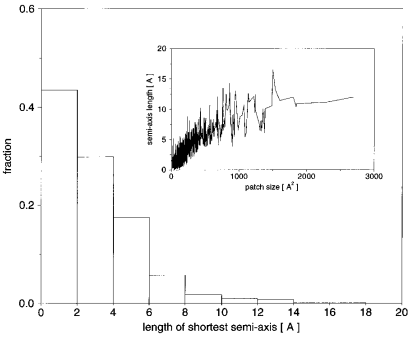

Fig. 12. Distribution of the lengths of the shortest semiaxis of the ellipsoids fitted to each interface patch as an indication of patch roughness. The inset shows this length as a function of the patch size.

occurrence as the patches become progressively more elongated.

The ellipsoids are generally oblique (disklike) as judged from their shortest semiaxis, *c*, substantially smaller than both *a* and *b* (Fig. 12). This is expected, given that they are fitted to a layer of solvent-accessible atoms. The length of semiaxis *c* provides a measure for patch "bumpiness." It is related to the root of the mean squared (RMS) distances of patch atom centers from the plane through the patch containing the *a* and *b* semiaxes. On average, the RMS distance is roughly 2.12 times the length of the semiaxis *c* for each patch. For large patches, the large-scale curvature of the protein will become noticeable in the measure for patch bumpiness. This effect can be seen in the inset to Figure 12, which shows a correlation of patch size with length of the shortest semiaxis. There are fewer occurrences of bumpy patches than of smooth ones as a result of the lower frequency of large patches. The results concerning patch form and bumpiness are indistinguishable from those found on monomeric protein surfaces (data not shown).

Composition

The amino acid composition of subunit interface patches was defined as the summed interface-accessible surface area of atoms of a given residue type contributing all patches of a certain size class divided by the total interface surface area of all patches in that size class. The results are shown in Figure 13. The largest contributors are the aliphatic and aromatic amino acids as well as proline. Leu, Ile, Phe, Val, and Pro occur progressively more often in larger patches, whereas the contributions of Trp and Tyr are roughly independent of the patch size. Ala and Met are intermediate contributors, the former as a consequence of its relative abundance, and the latter ensuing from its size and preference for interfaces.[2] The charged amino acids contribute to interface patches, but less so, and especially as the patch size grows. The Lys fraction is nonetheless considerable owing to the large apolar portion of its side chain.

There are some marked differences with the results found for monomeric protein surface patches (data not shown). The trends reflect the difference in composition between monomeric and interface surfaces. The contributions of Leu, Ile, Val, Phe, Tyr, and Met are higher in the interface patches than in the monomers with Lys, Asp, and Glu less common. The latter two show a more pronounced decrease with increasing patch size than in the monomeric case.

The differences between interface and exterior surface patches become clearer by taking the ratio of the surface composition for interfaces of a residue type to its composition on the exterior surface of protein multimeric complexes. If such propensity values are greater than or less than 1.0, the amino acid type is, respectively, preferred or avoided in the interface patches. The trends are shown in Figure 14. The aliphatic and particularly the aromatic amino acids are seen to prefer the interface patches, especially the larger patches. With the exception of Cys, the smaller polar residues do not display a strong preference for either environment. The charged residues, with the exception of Arg, show a definite dislike for interface patches, which is independent of their size. The preferences obtained here

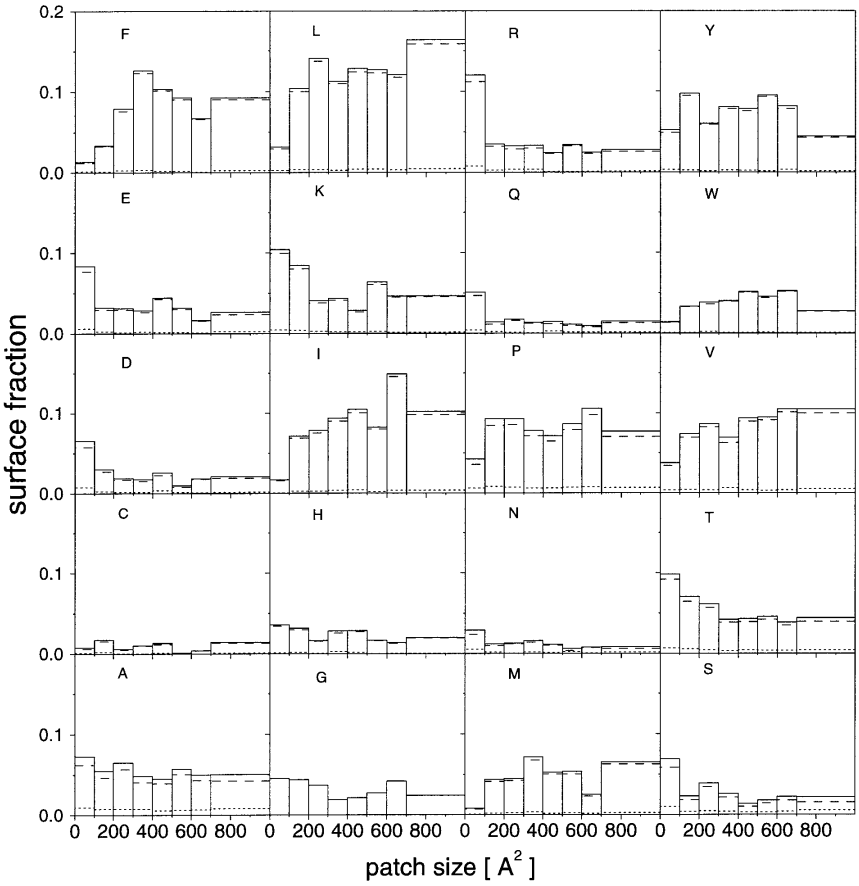

Fig. 13. The fractional surface contribution of amino acid types to interface hydrophobic patches according to their size. Values given are averages over the particular size range. All patches larger than 700 Å^2 were placed into the 700-and-greater bin. Solid lines indicate surface contributions from main- and side-chain atoms, while dotted and dashed lines show respective results for main chain only and side chain only atoms.

are in general agreement with those found for subunit interfaces by Argos.[2]

CONCLUSIONS

A survey of hydrophobic patches on the interfaces of protein subunits has shown many consistent characteristics. There are more large patches on interfaces than on protein exteriors. The distribution of the sizes of the largest patch in each subunit interface peaks at about 500 Å^2, albeit one patch displayed 2600 Å^2; and yet there is no correlation between the size of the largest patch and the size of the interface surface. The size distribution of the largest interface patch strongly resembles that for monomeric protein surfaces. If a hydrophobic patch displays any overlap with the interface, this overlap tends to be large, which could well be significant in the process of subunit association and recognition.

There is a very high coincidence between the largest patches on the exterior surface of a subunit

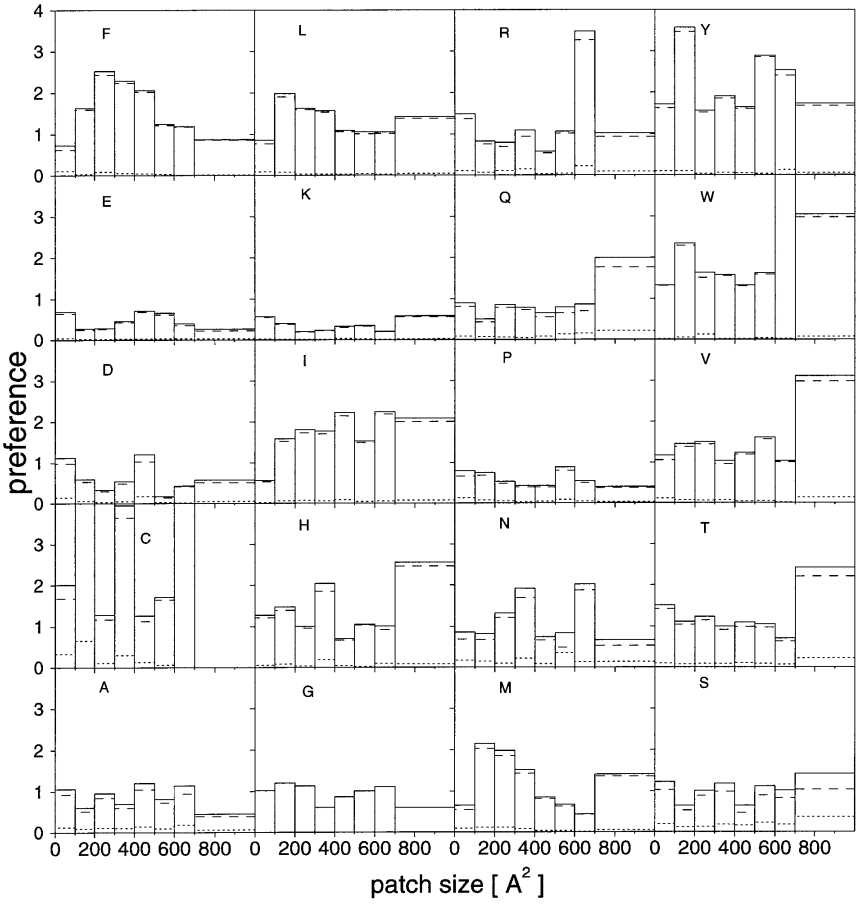

Fig. 14. The surface preference of amino acid types to be within interface patches relative to their contribution to exterior surface hydrophobic patches on the multimeric complexes. Values given are averages over the particular size range. All patches larger than 700 Å were placed into the 700-and-greater bin. Solid lines indicate preferences when considering surface contributions from main- and side-chain atoms, while dotted and dashed lines show respective results for contributions from main chain only and side chain only atoms.

and the interface: in 90% of the proteins in our sample, either the largest or the second largest constituted a significant portion of the interface. This should prove valuable in studies of protein–protein association. If the subunit association site of a polypeptide chain is not known, the largest or second largest patch is the most likely candidate. They are also the best mutagenesis targets if oligomerization potential is to be altered. Further, if the tertiary structure of a subunit is known, the two largest hydrophobic patches would greatly limit the search space for docking oligomers.

The complementarity of contacts of patches on different subunits was found to be low in the sense that large patches on one subunit do not often mostly cover large patches on the associating subunit. The

shielding of hydrophobic surface from water appears more important than contacts between similarly sized apolar surface patches. The relative contributions of amino acid types and their composition on patches of different sizes reflects their general composition on interfaces.

ACKNOWLEDGMENTS

The authors thank Jaap Heringa for helpful discussions and Nelly van der Jagt for invaluable assistance in the preparation of the manuscript.

REFERENCES

1. Janin, J., Miller, S., Chothia, C. Surface, subunit interfaces and interior of oligomeric proteins. J. Mol. Biol. 204:155–164, 1988.
2. Argos, P. An investigation of domain and subunit interfaces. Protein Eng. 2:101–113, 1988.
3. Korn, A.P., Burnett, R.M. Distribution and complementarity of hydrophobicity in multi-subunit proteins. Proteins 9:37–55, 1991.
4. Jones, S., Thornton, J.M. Protein–protein interactions: A review of protein dimer structures. Prog. Biophys. Mol. Biol. 63:31–65, 1995.
5. Lijnzaad, P., Berendsen, H.J.C., Argos, P. A method for detecting hydrophobic patches on protein surfaces. Proteins 26:192–203, 1996.
6. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F. Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. The Protein Data Bank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
7. Abola, E.E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F., Weng, J. Protein Data Bank. In "Crystallographic Databases: Information Content, Software Systems, Scientific Applications." Alle, F.H., Bergerhoff, G., Siebers, R. (eds.). Bonn/Cambridge/Chester: Data Commission of the International Union of Crystallography. 1987:107–132.
8. Etzold, T., Argos, P. Transforming a set of biological flat file libraries to a fast access network. Comp. Appl. Biosci. (CABIOS) 9:59–64, 1993.
9. Etzold, T., Argos, P. SRS: An indexing and retrieval tool for flat file data libraries. Comp. Appl. Biosci. (CABIOS) 9:49–57, 1993.
10. Heringa, J., Sommerfeldt, H., Higgins, D., Argos, P. OBSTRUCT: A program to obtain largest cliques from a protein sequence set according to structural resolution and sequence similarity. Comp. Appl. Biosci. (CABIOS) 8:599–600, 1992.
11. Eisenhaber, F., Lijnzaad, P., Argos, P., Sander, C., Scharf, M. The double cubic lattice method: An efficient approach to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies. J. Comp. Chem. 16:273–284, 1995.
12. Lijnzaad, P., Berendsen, H.J.C., Argos, P. Hydrophobic patches on the surfaces of protein structures. Proteins 25:389–397, 1996.
13. Young, L., Jernigan, R.L., Covell, D.G. A role for surface hydrophobicity in protein–protein recognition. Protein Sci. 3:717–729, 1994.
14. Taylor, W.R., Thornton, J.M., Turnell, W.G. An ellipsoidal approximation of protein shape. J. Mol. Graphics 1:30–38, 1983.

6

PATCH DYNAMICS

P. Lijnzaad, K. A. Feenstra,
J. Heringa and F.C.P. Holstege
On Defining the Dynamics of Hydrophobic Patches on Protein Surfaces.
Submitted to Proteins

ABSTRACT

We present a simple and efficient method called PATCHTRACK, for studying the simulated dynamics of hydrophobic surface patches. It connects static patches on snapshot structures through time into so-called patch runs, which are subsequently clustered into so-called recurrent pat-ches. The method is applied to simulations of three different proteins. Protein motion causes addition and removal of one or more atoms to a patch, resulting in size fluctuations of around 25%. The fluctuations eventually lead to the break-up of a patch, and their average life span is therefore remarkably short at around 4 ps. However, some patch runs are much more stable, lasting hundreds of picoseconds. One such case is the largest patch in amicyanin that is known to be biologically relevant. Another case, previously not reported, is found in phospholipase A_2, where the functional significance of a large recurrent patch formed by Leu58 and Phe94 appears likely. The most frequently occurring patch size is 40-60 $Å^2$, but sizes of up to 500 $Å^2$ are also observed. There is no clear relation between patch run durations and their average size. However, long-lasting patch runs tend not to have large fluctuations. Although the algorithm would allow the patches to "wander" over the surface, this does not happen in practice. The recurrent patches have alternating periods of "liveness" and "dormancy"; around 25% of them is predominantly in the live state.

INTRODUCTION

Hydrophobic patches on protein surfaces are crucial in protein-protein binding,[1, 2] protein-ligand binding (*e.g.* Refs 3, 4), and protein folding.[5, 6] Hydrophobic interactions are often functionally significant, but they may also be inadvertent: a number of pathologies are explained by such hydrophobic aggregation.[7]

It has long been realized that the mobility of parts or all of the protein structure is essential in protein function.[8, 9] Given the importance of protein structural dynamics, it is interesting to probe the dynamics of hydrophobic patches using Molecular Dynamics simulations. However, the dynamics of hydrophobic patches have, to our knowledge, never been defined well, nor studied in detail.

In order to explore the dynamics of patches, a definition of hydrophobic patches is required. We use the QUILT method proposed in earlier work,[10] because it is, to our knowledge, the only method that yields well defined hydrophobic patches in atomic detail.

The current work describes, firstly, how the QUILT patches can be used to study the patch dynamics. Secondly, their general trends in simulations of proteins are established, to serve as a baseline against which to judge the behaviour of hydrophobic patches in the in-depth studies of protein-ligand dynamics.

METHODS

PROTEIN STRUCTURES

We chose three different proteins, spanning a reasonable range of sizes and architectures: the immunoglobulin G-binding domain B1 of streptococcal protein G (PDB code **1pgb**); amicyanin (PDB code: **1aaj**), and the closed, inactive form of phospholipase A_2 (PDB code: **1p2p**). Of the three structures, only amicyanin (**1aaj**) has a known functional hydrophobic patch, which is involved in binding to methylamine dehydrogenase during electron transfer.[11] For brevity, the proteins are referred to by their PDB accession code in the remainder of the text. The protein structures used in this work were taken from the Protein Data Bank. Some essential details of these structures are given in Table 1.

	protein G Ig-binding domain B1	**Amicyanin**	**Phospholipase A_2**
PDB code	**1pgb**	**1aaj**	**1p2p**
Reference	12	13	14
Resolution [Å]	1.92	1.8	2.6
Residues	56	105	124
Heavy atoms	436	807	971
Crystallographic waters	24	98	5
*Waters added**	2168	2937	5025
*SAS [Å^2]***			
X-ray	3668	5454	7089
MD†	3802 $\pm$ 72	5848 $\pm$ 72	7653 $\pm$ 103
Percentage hydrophobic			
X-ray	58	62	60
MD†	59 $\pm$ 0.9	61 $\pm$ 0.8	58 $\pm$ 0.8
Number of patches‡			
X-ray	13	21	31
MD†	16.9 $\pm$ 1.9	26.9 $\pm$ 2.5	28.5 $\pm$ 2.5
min;max	10; 23	19; 35	19; 36

*for simulation purposes
**Solvent accessible surface area; probe radius 1.4 Å
†average $\pm$ standard deviation, over 300 ps simulation
‡QUILT patches larger than 10 Å^2

Table 1: *Details of the protein structures.*

SIMULATIONS

All simulations were carried out with the GROMACS package.[15] The simulation protocol was as follows: atomic coordinates were extracted from the PDB file, and a molecular topology was generated. Periodic boundary conditions were employed, using a rhombic dodecahedron cell with a separation of at least 8 Å between protein atoms and the nearest side of the cell. Crystallographically resolved water molecules were kept, and the system was immersed in a box of pre-equilibrated SPC water molecules[16] at 300 K. For both **1pgb** and **1aaj**, the system had a net charge of $4e^-$, so in both cases, 4 water molecules with high electrostatic potential were replaced by Na^+ ions. The half-occupied Ca^{2+} site in the **1p2p** structure was left out to achieve electric neutrality. The 53A6 forcefield[17] was used.

Bond lengths were constrained with the Lincs algorithm.[18] The time step of the integrator was 2 fs. Neighbour lists for the calculation of non-bonded interactions were recalculated every 10 time steps, and a twin-range cut-off radius was used for evaluating them. Within 9 Å, they were calculated every time step, whereas those within 12 Å were only evaluated every 10 time steps. Temperature was kept constant at 300 K by coupling to a thermal bath with coupling constant τ_T=0.1 ps; pressure was kept constant by coupling with constant τ_P=0.5 ps to atmospheric pressure, 10^5 Pa.

The system's energy was minimized until convergence to machine precision using the steepest decent optimizer, and the water was subsequently equilibrated by restraining the protein atoms to their initial positions while integrating the system for 10 ps at 300 K. All restraints were removed, and after 20 ps of Molecular Dynamics integration to achieve equilibrium, a 300 ps trajectory was obtained. From these trajectories, snapshots taken at 0.1 ps intervals were used in the subsequent analyses.

PATCH DETECTION

The analysis of the dynamics of the patches requires the following ingredients: the detection of patches on each static structure using QUILT; tracking each QUILT patch over time, by assembling them into so-called *patch runs*; and lastly, assembling patch runs into so-called *recurrent patches*.

A brief overview of QUILT[10] is as follows. The solvent accessible surface of a protein is considered a contiguous surface of adjacent hydrophobic (carbon and sulfur) atoms, surrounded by strings of hydrophilic (nitrogen and oxygen) atoms. Most proteins surfaces are relatively apolar (around 60% by solvent accessible surface area), so the connection of neighbouring solvent-accessible parts of hydrophobic atoms would result in one large surface patch spanning the entire protein. This surface is dotted with polar "islands" formed by the hydrophilic atoms, with hydrophobic connections through variously sized "channels" between between these islands. To delineate the hydrophobic patches, the channels are closed off by

temporarily expanding all solvent-accessible polar atoms by a fixed "polar expansion radius". Thus, the surface is divided into isolated proper patches. They are enumerated, and adjacent surface area lost due to the polar expansion is added back.

For QUILT, the default probe and polar expansion radius (1.4 Å) were used.

TRACKING PATCHES OVER TIME

Given QUILT's patch definition, a particular patch on the protein surface will change over time: it changes size and shape, it may loose or gain atoms (typically at the edges), it may split or merge; it may even disappear and re-emerge altogether. That is, a patch is a dynamic entity with varying constituents. To study the dynamic behaviour of patches, the terminology has to be made more precise. In the current context, a QUILT patch is the occurrence of a patch on one snapshot of a structure taken from the Molecular Dynamics trajectory. We will call this a *snapshot patch*.

A snapshot patch may be followed across subsequent time frames of a trajectory. We will call one such run of connected snapshot patches a *patch run*. Establishing them is the second step of the procedure.

As will be discussed below, patch runs are generally limited in duration, as protein motion causes changes in their size, eventually leading to their break-up. Although patch runs do not last very long, they often re-emerge a while after their disappearance. We will call such a series of patch runs a *recurrent patch*, and they are obtained in the last step of the procedure.

IDENTIFICATION OF PATCH RUNS

The method used to assemble snapshot patches into patch runs uses no geometrical criteria. Instead, it is based solely on the identities of the atoms constituting the snapshot patches. By optimizing the matching between all QUILT patches on both snapshot structures, correspondences between snapshot patches on either structure are obtained. This allows the connection of each snapshot patch with its predecessor and successor, yielding a chain of snapshot patches that constitute a patch run.

To judge the matching of patches on two structures, we measure the overlap between two patches p and q on structures A and B using the so-called *Jaccard coefficient*[19] of the two sets of atoms p and q: $J_{AB}(p,q) = \frac{\|p \cap q\|}{\|p \cup q\|}$. That is, the overlap coefficient is defined as the number of atoms present in both snapshot patches, divided by the number of atoms in their union. $J_{AB}(p,q)$ is 1 when p and q are identical, and 0 when they are fully disjunct.

For speed, the lists of snapshot patches p and q are each ordered by decreasing size, and for each of the snapshot patches p, the overlap with all snapshot patches q is calculated. The pair (p_i, q_j) for which $J_{AB}(p_i, q_j)$ is maximum, is declared a match; p_i and q_j are removed from both lists, and the process is reiterated until either list of p_i's or q_i's is empty (*i.e.*, this is a greedy algorithm).

To avoid poor matches causing spurious matchings, an overlap cut-off value J_{rej} is applied below which a match (p_i, q_j) is rejected, even if it has the maximum overlap coefficient between p_i and any from the list of q patches. Without a cut-off, a poor $J_{AB}(p_m, q_j)$ early in the iteration over the p's occasionally rules out a better $J_{AB}(p_m, q_j)$ later in the iteration. Increasing the J_{rej} makes the matching process less greedy. This improves the overlap coefficient for matched snapshot patches, but it reduces the overall matching because the number of snapshot patches that fail to be matched altogether also rises. A J_{rej} between 3% and 12% was determined to be optimal (data not shown); we used 5% in this work.

IDENTIFICATION OF RECURRENT PATCHES

In the last step, patch runs are assembled into recurrent patches. As with the patch run identification described above, only atom identities are used. To this end, the set of "core atoms" of a patch run has to be defined. The obvious choice would be the set of atoms that is present in all snapshot patches of the patch run. We will call these atoms the *fully persistent atoms* of a patch run. For brevity, the number of fully persistent atoms will be called N_{fpa}. They are found by simply taking the intersection amongst all sets of atoms of all snapshot patches of a patch run. However, the resulting set of fully persistent atoms is frequently small ($N_{fpa} \leq 1$) as a result of fluctuations. The converse option is to use the union of the atoms in all snapshot patches of the patch run. This becomes unwieldy however, because these sets can become large. We take the middle ground, by defining the core atoms of a patch run as those that are present during at least half of its duration. The correlation of their number with the area of a patch run is greater than 0.75 in the three systems studied.

The next step relates the patch runs by clustering them based on their core atoms. The distance measure is the inverse of the formula used for the snapshot patch overlap: $D = \frac{1}{J(p,q)} = \frac{\|p \cup q\|}{\|p \cap q\|}$, with p and q the respective sets of core atoms of the patch runs being compared. Whenever the patch run overlap is zero, the distance was arbitrarily set to $f_{max} = 10$ times the maximum distance found amongst all pairs of patch runs, to avoid infinite values in the distance matrix. The data is clustered using group-average linkage (UPGMA), and cutting the dendrogram at an ultrametric distance $d_{cut} = 10$ produces tight clusters of patch runs. These clusters are the recurrent patches.

Unlike their constituent patch runs, the recurrent patches do not have a duration; they are features of the dynamical protein structure in general. They do however have alternating periods of presence and absence of the patch runs that constitute it. We will call these "states" of the recurrent patch *dormant* and *live*, respectively.

Programs and scripts (available from `http://www.bioinformatics.med.uu.nl/publications/lijnzaad/patchdynamics/`) were written in Perl,[20] R,[21] and standard Unix utilities. In addition to these scripts, use has been made of xmgr/grace,[22]

RasMol,[23] PyMol (DeLano Scientific LLC), and some of the programs in the GROMACS suite.[15]

RESULTS

MATCHING PROCEDURE

The degree of matching can be defined as the fraction of patches mapped between one snapshot and it successor. We find an average, over all time frames, of around 80%. The time interval Δt between the snapshots influences the degree of matching, as shown in Fig. 1. The data is obtained by averaging over matchings between all pairs of snapshots from a 10 ps interval from the trajectory of structure **1aaj**.

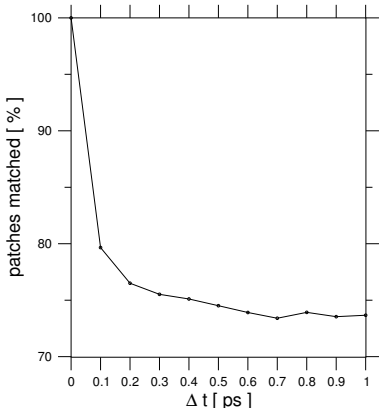

Figure 1: *Percentage of patches matched between two time frames as a function of the time step Δt between them.*

PATCH RUNS

Visual inspections of the patch runs using movies of the molecular trajectory revealed that the matching behaves as expected: a snapshot patch in one frame looks largely the same as its match in the next frame, providing a smooth dynamic picture of its behaviour.

An impression of how the size of a patch evolves during the patch run is given in Fig. 2, which shows the surface area of the three longest-lasting patch runs in each simulation. The longest-lasting of all patch runs lasts around 250 ps; it is also the largest amongst all patch runs. It was found in the simulation of **1aaj**, and

corresponds to the hydrophobic patch known to be involved in the interaction of this protein with methylamine dehydrogenase.

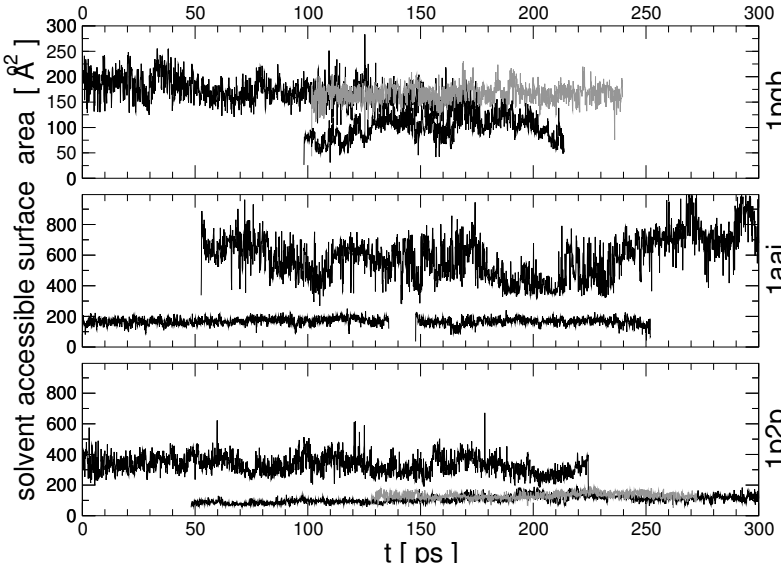

Figure 2: *The surface area of the three longest-lasting patch runs in each simulation. The color grey is used if curves overlap. PDB codes of the structures in each row are indicated on the right.*

As mentioned in *Methods*, the number of patch runs with an appreciable set of fully recurrent atoms ($N_{fpa} > 1$) is generally low (**1pgb**: 28%; **1aaj**: 25%; **1p2p**: 32%).

An important characteristics of a patch run is its "lifespan", called its *duration*. Since there is little to conclude about the "real" duration of a patch run that is already existent when the simulation starts, or is still ongoing when it ends, patch runs crossing the temporal boundaries of the simulation were discarded, as were those lasting shorter than 1 ps.

We define the size of a patch run simply as the average area of its constituting snapshot patches; their standard deviation will be called its *fluctuation*. Statistics (*e.g.*, average *etc*) and distributions on the size and fluctuation of a patch run (used below) ignore the fact that they themselves are also defined as an average. Some statistics on the resulting patch runs are given in Table 2.

PDB code	**1pgb**	**1aaj**	**1p2p**
Number of patch runs	788	1479	1365
Duration [ps]			
average	4.0	3.7	3.9
standard dev.	9.1	5.9	7.3
median	1.9	1.9	1.7
maximum	145	104	138
Size [Å^2]			
average	97	89	108
standard dev.	69	47	78
minimum	12	3	6
median	81	77	77
maximum	484	320	426

Table 2: *Statistics on the patch runs. Those lasting shorter than 1 ps or existing at $t = 0$ or at $t = 300$ ps are excluded.*

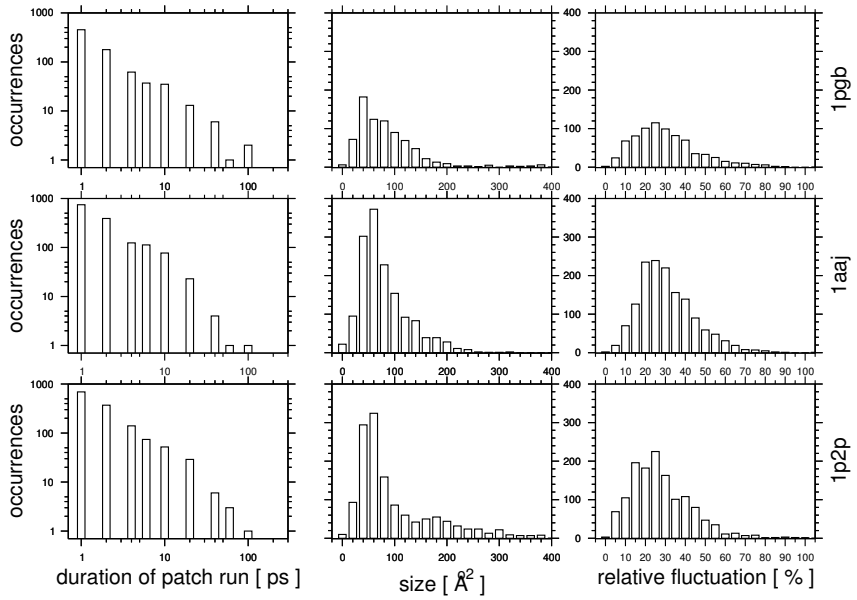

Figure 3: *Distributions of patch run duration, size, and size fluctuations. The columns depict the distributions of the duration of patch runs (left), their size (middle) , and their fluctuation (right). The data in the left column is binned and plotted logarithmically. Those in the middle and right columns contain linear plots, with vertical axes at the same scale. PDB codes of the structures in each row are indicated on the right.*

The average patch run lasts around 4 ps, although a substantial number is more persistent, exceeding 100 ps. Their mean size is around 90 Å^2, but the

maximum patch run sizes are a few hundred square Ångstroms.

Fig. 3 shows the distributions of duration, size and fluctuation in more detail. It shows that short-lived patch runs are much more prevalent than longer-lasting ones. This trend is so strong that the only meaningful presentation of the distribution is as a log-log plot. The distributions of patch run sizes shows that the most frequently occurring size is 40-60 Å^2. The fluctuations depicted in Fig. 3 are shown as a relative value (standard deviation divided by mean, also known as the coefficient of variation) to avoid trends being obscured by the mass of small patches having a low absolute fluctuation. The most frequently occurring value for the relative fluctuations is around 25%.

SIZE, FLUCTUATION AND DURATION

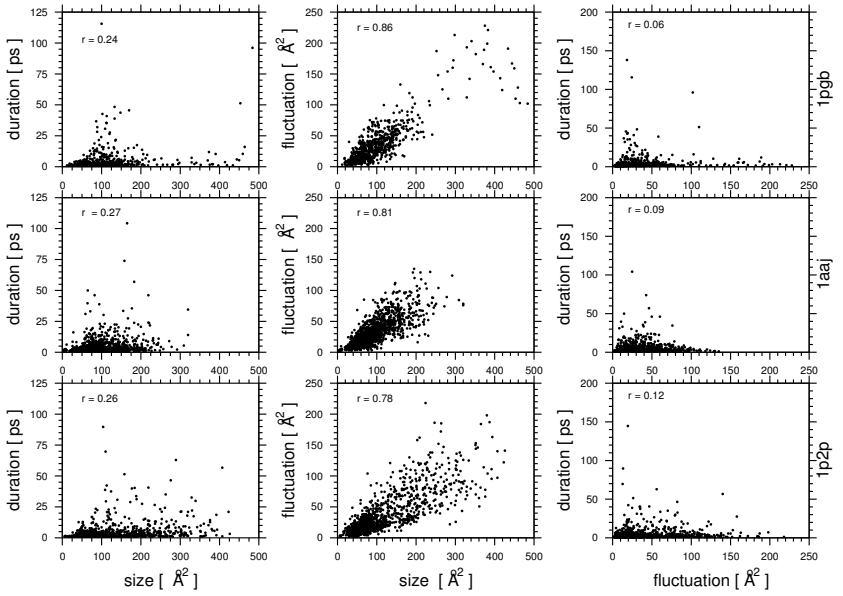

Figure 4: *Scatter plots of the patch run size and duration (left), size and fluctuation (middle) and fluctuations and duration (right). Linear correlation coefficients are indicated in each graph. PDB codes of the structures in each row are indicated on the right.*

Fig. 4 shows scatter plots of the average size of patch runs, their life spans, and fluctuations.

There is hardly any relation between duration and size. Only some mid-sized patch runs reach long life-spans; large patches do not, although there are a few exceptions. There is a roughly linear relationship between the average size of a patch run and its fluctuations. The last column of scatter plots in Fig. 4 depicts

the relation between fluctuation and duration of a patch. Its correlation is even lower than that between duration and size. We also looked into the following patch characteristics:

- fraction of hydrophobic amino acids
- fraction of backbone atoms
- fraction of sequentially remote amino acids
- shape (ratio of the largest two semi-axes of a fitted ellipsoid)

None of these correlated to an appreciable extent with duration, size or fluctuation (the highest correlation observed was 0.42).

RECURRENT PATCHES

To obtain the recurrent patches, the patch runs are clustered and the resulting tree is cut at an ultrametric distance of 10, as described under *Methods*. Visual inspection showed that this works well. The cut-height is not critical, because the clusters are very tight, and the distance to the next clustering level is large (data not shown). This is also clear from the Silhouette width, a measure of cluster compactness.[24] Its value is around 0.7, averaged over all clusters obtained by cutting the tree, and depends little on the choice of the cut-height. The clustering method is also not critical; group-mean average, complete, median, centroid and Ward's method all give similarly compact clusters.

To get an impression of how patch runs appear and disappear in the recurrent patches, they are plotted in Fig. 5. It shows how recurrent patches alternate between the "dormant" and "live" states. The latter are indicated by the blocks in the figure. Most recurrent patches are dormant most of the time; few are predominantly in the live state. For many recurrent patches, some of their patch runs overlap briefly in time (data not shown). On average, such temporal overlaps occur during approximately 3% of the simulation time.

For most recurrent patches, the coefficient of variation of the size of their constituting patch runs comes out at around 25%, the same as that found for the patch runs.

We define the *liveness* of a recurrent patch as the percentage of simulation time during which it is in the live state. Statistics on the liveness of recurrent patches are given in Table 3. As mentioned, few recurrent patches are mostly live. Therefore, the table also shows the statistics for a high-liveness ($\geq$ 50% live) and a low-liveness ($<$ 50% live) group.

The number of recurrent patches is somewhat larger than the number of QUILT patches in the static structure of a protein (see Table 1). The high-liveness group is 2 to 3 times smaller than the low-liveness group. The contrast in the liveness of these groups is quite stark: the high-liveness recurrent patches are 5 times as often in the live state.

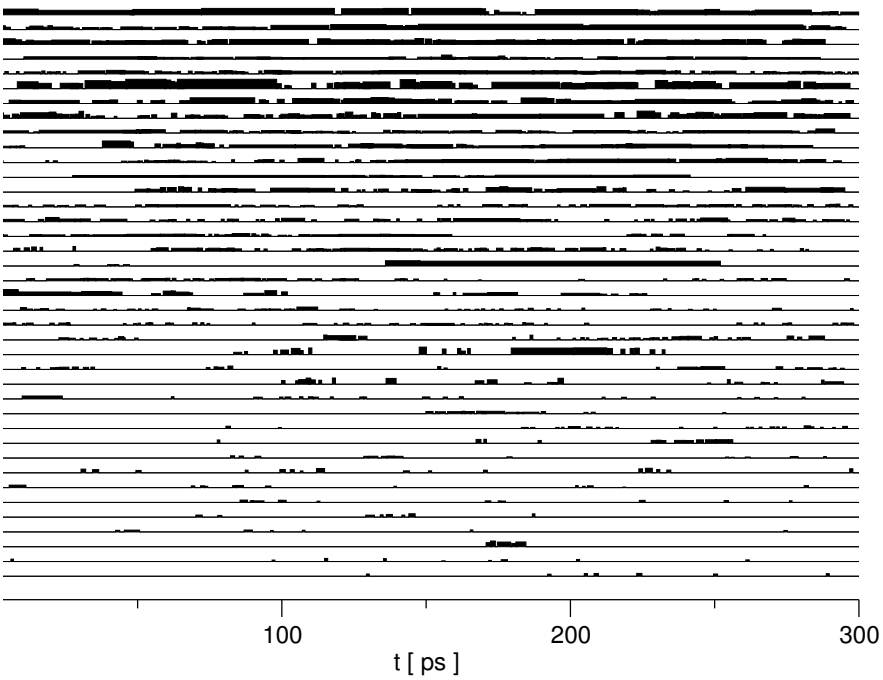

Figure 5: *Recurrent patches on structure **1aaj**. Each track represents one recurrent patch. Each block is one patch run, with thickness proportional to its average size. The tracks are in order of liveness; the vertical separation is 500 Å^2. Recurrent patches with an average area below 50 Å^2 or consisting of one patch run are excluded.*

PDB code	**1pgb**	**1aaj**	**1p2p**
Number of recurrent patches	35	55	63
liveness $\geq$ *50%*	9	16	13
liveness $<$ *50%*	26	39	40
Liveness [%]			
average	29	32	26
$\geq$ *50%*	78	77	75
$<$ *50%**	12	13	14
minimum	0.4	0.3	0.4
maximum	94	96	96

*The subset-averages need not sum to unity as the liveness distribution is asymmetric

Table 3: *Statistics of recurrent patches. Overall statistics are given, as well as those obtained by first separating the recurrent patches into a "high-liveness" ($\geq$ 50% live) and "low-liveness" ($<$ 50% live) group.*

DISCUSSION

MATCHING PROCEDURE

The matching procedure works well, based on visual inspection and as judged from the fraction of the number of patches matched (around 80%). The snapshot patches tend to be of similar size during the course of the patch run. This is achieved by the use of the Jaccard coefficient. For instance, it classifies the containment of a small snapshot patch by a large one as poor, because the large snapshot patch still contributes to the denominator of the coefficient. As a result, the case of one snapshot patch p_i simultaneously containing two smaller patches q_j and q_k is dealt with correctly, because the larger of the two q's will have the higher overlap coefficient.

Visual inspection of patch runs showed that the patches do not "wander" over the protein surface. This is confirmed by the clustering, into recurrent patches, of patch runs that are widely separated in time (see Fig. 5). The PATCHTRACK algorithm would, in fact, accommodate and therefore allow such wandering of patches. The fact that this does not occur suggests that they are genuine protein surface features.

In general, wandering is not expected, because sites that interact with ligands or other proteins tend to be fixed. One can, however, conceive of a situation in which such wandering would be functional. For instance, on the surface of a filamentous protein, the concerted motion of groups of hydrophobic residues could result in a hydrophobic patch moving along the filament, serving as a "conveyor belt" for hydrophobic compounds. The PATCHTRACK formalism would be able to cope with such a case, although it does not apply to the proteins studied here.

As the time interval Δt between two snapshot structures increases, the root mean squared deviation (RMSD) between the two structures, especially that of solvent accessible atoms, increases likewise. Consequently, the matching becomes more difficult as can be seen in Fig. 1. It shows that a sampling frequency of less than once per 0.1 ps results in a considerably lower degree of matching. Conversely, sampling at a higher frequency should increase the degree of matching, but this is rarely used in practice.

Theoretically, following patches can be done in a geometrical fashion, for instance by clustering the coordinates of the patch centers. This was attempted initially, but it is slow due to the large amount of data, and very sensitive to the choice of parameters. A more serious drawback is that geometric matching may not be possible, or be very difficult to parameterize, if there are large scale changes in the proteins structure over the course of the simulation. PATCHTRACK deals with this automatically by using atom identities only.

QUILT uses two parameters: the probe radius and polar expansion radius. The PATCHTRACK method introduces three more: the patch overlap cut-off J_{rej} for the matching procedure, and, for the clustering step, the maximum distance

multiplication factor f_{max} and the cut-height d_{cut}. Their default values should not need adjustment in the large majority of cases, as results are not sensitive to the exact settings. The matching and clustering procedures perform comparably for the three different protein structures, indicating that the method is robust.

PATCH RUNS

The fluctuations in patch size can be substantial, as shown in Fig. 2. They occur at different time scales, small fluctuations lasting tenths of picoseconds, superimposed on larger fluctuations lasting up to 100 times longer. This reflects the hierarchy in the motion of atoms, side-chains and secondary structure elements.[25, 26]

Many natural entities (*e.g.*, populations) behave gradually: they start small, grow in size steadily, persist for a while, shrink slowly, then cease existing. This is not the case for patch runs; they start and stop fairly abruptly. Although the areas of the snapshot patches at either end of the life span is a bit below average, they are by no means atypical, given the large fluctuations displayed during the patch run. However, the fluctuations do, of course, cause the end of patch runs, in that it is matter of time until the fluctuation is so large that it destroys the patch altogether.

There are two mechanisms behind the fluctuations and the volatility of the patch run. When an atom joins a patch, it can easily add 20-30 Å^2 to its area. With average patch sizes of around 100 Å^2 (*cf* Table 2), such additions (and likewise, removals) are substantial. This is also visible in the distribution of fluctuations shown in Fig. 3, which shows that a relative fluctuation of around 25% is the most commonly occurring one. Atoms can join and leave patches due to brief burials, but more importantly, due to the fact that they frequently alternate between neighbouring patches. This, in turn, is caused by small shifts in the positions of the polar atoms used by QUILT to delineate patches. It is then a matter of time until so many atoms are by chance simultaneously absent from the patch under consideration, that it has become too small to be matched properly to the corresponding snapshot patch in the next time frame. The result is the end of the patch run. The shifts of the atoms leading to reassignment happen at the sub-picosecond time scale, explaining the time scale of the fluctuations.

A second, for large patches more important mechanism generating large fluctuations is a scaled-up version of the first mechanism: the addition or removal of a complete group of atoms to the patch under consideration. Typically, the atoms in this group belong to one amino-acid side chain. An example of this is shown in Fig. 6. It depicts the solvent accessible surface of structure **1aaj** at time points 174.2, 174.3 and 174.4 ps. Over these 200 femtoseconds, the patch area plummets from 945, to 567, to 400 Å^2. This is the result of the removal of both Phe97 and Lys27 in the first time step, and Lys73 in the second time step. The overall changes in the structure are minimal, but they conspire to reduce the surface of this patch by more than half.

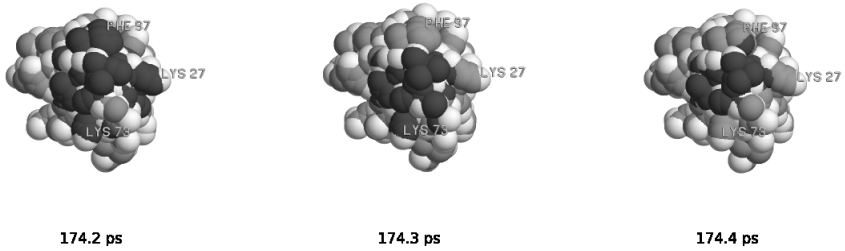

Figure 6: *Three successive time frames from the **1aaj** simulation. The figure shows the solvent accessible surface area of the complete protein. White: oxygen and nitrogen atoms; light-grey: carbon and sulfur atoms; dark-grey: atoms belonging to largest and longest-lasting patch. Time points and surface area of the snapshot patch are indicated below the snapshots.*

Looking at the set of atoms contributing to a patch run over its life time, we find similar volatility, as few patch runs have many fully persistent atoms. N_{fpa} is not correlated with the patch run size, nor with patch run duration (data not shown).

The long-lasting patch runs shown in Fig. 2 last between 100 and 250 ps. If longer simulation times are used, the maximum duration will probably be greater, especially so for the patches already existing at $t = 0$ or still existing at $t = 300$. The long-lasting patches have sizes well above the average (*cf* Table 2), but they are mostly not the largest patches. The exceptions are those in the **1aaj** (amicyanin) and **1p2p** (phospholipase A_2) simulations, at around 600 and 400 Å^2, respectively. They are both the longest-lasting and largest patches in their simulation. The large recurrent patch on the structure of amicyanin is known to interact with methylamine dehydrogenase.[11] The large recurrent patch on phospholipase A_2 is not known to be functionally significant, but may well be so (see below). If this is indeed the case, then these two examples suggest a hypothesis for further enquiry, namely that functional patches tend to be not only large,[10] but also persistent.

The graph for amicyanin (Fig. 2; **1aaj**) shows two smaller patch runs with an area fluctuating around 200 Å^2, separated by around 10 ps. The atoms making up these patch runs are nearly identical, and this is a clear example of two patch runs that are part of one recurrent patch.

None of the patch runs exceeds our 300 ps simulation time (Fig. 2), showing that the relatively short simulation times used in this study are sufficient to capture the essentials of the dynamics of the QUILT patches.

The average patch run duration is about 4 ps, a time scale similar to that found for simulated[27–29] and experimental[30–32] residence times of water molecules in the first hydration shell of proteins.

There is generally no relation between the duration and size of a patch run, as shown in Fig. 4. We conclude that the duration and size of a patch run are so

dependent on the details of the local mobility that no general trend emerges.

In general, larger quantities usually display larger fluctuations, and this is also observed for patch runs and their fluctuations. It is interesting to see that this relation is roughly linear, as one would expect a patch to grow or shrink predominantly at its perimeter. The area changes must then be proportional to the length of the perimeter, and therefore proportional to the square root of the patch area. This, however, ignores the roughness of the patch perimeter. If this would be taken into account, it is likely that the perimeter as a function of area varies with an exponent closer to 1, explaining the apparent linear relationship between patch size and fluctuation.

There is no linear correlation between duration and fluctuation (Fig. 4), but there is an obvious relationship, namely, that long-lasting patch runs rarely occur in patch runs that exhibit high fluctuations. This is not surprising: the chances of a patch run surviving for hundreds of picoseconds are slim if the magnitude of the fluctuations increases the chances of disruption.

RECURRENT PATCHES

The recurrent patches were subdivided by their average liveness, using a liveness of 50% as the dividing line. This results in a relatively small (20-30%) group of recurrent patches with strong liveness ($\geq$ 75%; Table 3). The large remaining group is predominantly dormant, and consist of relatively small and brief patch runs caused by splits from their "main" patch. They are the inevitable "noise" inherent in a patch detection method based on contiguous atomic surface area, and are best ignored when focussing on recurrent patches. Because of the small number of highly live recurrent patches in the current work, we have not attempted to relate the liveness to the size or duration of the constituting patch runs.

A NOVEL PHOSPHOLIPASE PATCH

The large, persistent patch on phospholipase A_2, visible in Fig. 2, is an interesting case. In the crystal structure, it is much smaller (220 Å^2) than in the simulation (400 Å^2). It is formed by Leu58 at the C-terminal end of helix C, and Phe94 at the C-terminal end of helix E, at the "back" of the protein. It is far removed from the active site, and is not part of the ligand binding cleft. The patch could be a by-product of an interaction, between said residues, that stabilizes the two helices. However, this would seem unnecessary, as the helices are already tethered by a total of three disulfide bridges. The hydrophobic nature of the two residues at this position appears to be conserved across much of the Pfam[33] alignment (accession PF00068, Phospholipase A_2). Based on this and on the fact that the patch is both large and persistent, it may well be biologically relevant. Phospholipases require association with membranes or micelles for their activation,[34] but this interaction is ionic in nature and takes place at the other side of the molecule. We therefore speculate that the Leu58,Phe94 patch is important for a different reason, such as

association with lipophilic moieties of membrane compounds, or the stabilization of the short D helix in the (unknown) membrane-associated form of the protein. Elucidating the function will require experimental work, but the case demonstrates the value of the PATCHTRACK approach in generating hypotheses.

CONCLUSION

PATCHTRACK is a robust, fast procedure to follow hydrophobic patches on Molecular Dynamics trajectories of proteins. It connects the snapshot patches given by QUILT[10] into patch runs using a simple matching procedure based on atom identities. The patch runs are subsequently clustered into recurrent patches, again using only atom identities.

The patches do not "wander" over the protein surface, but the set of atoms constituting a patch run is not very stable, owing to protein mobility. Protein motion causes small fluctuations in patch size, whereas large fluctuations arise from the addition or removal of groups of atoms, typically from the same residue. The most commonly observed mean size of patch runs is 40-60 $Å^2$, but they can be much larger, up to a few hundred $Å^2$. The patch runs can last up to approximately 150 ps, but are generally much shorter; the average patch run lasts around 4 ps. Size fluctuations during a patch run are roughly proportional to the patch size. They are substantial, at around 25% of the patch size. The relation between patch run duration and the average size is not clear, but that between patch size fluctuations and their duration *is*: large fluctuations preclude long durations.

Clustering the patch runs into recurrent patches uncovers the really persistent patches. These are relatively few, as on average, the recurrent patches are in the live state during 25% of the time. However, the 50% most "live" patches have an average liveness of around 75%.

Amongst the persistent large patches in the systems studied, two stand out. The first is the well-known patch on amicyanin that is involved in the interaction with methylamine dehydrogenase. It has a size fluctuating around 600 $Å^2$, and lasts for 250 ps. Secondly, in phospholipase A_2, a novel patch that is probably functionally important is found. It consists of Leu58 and Phe94, measures around 400 $Å^2$ and lasts for over 220 ps.

In addition to QUILT's probe radius and polar expansion radius, PATCHTRACK introduces three parameters: the cut-off value for patch overlaps, and the maximum distance and cut-height for clustering. These parameters are not critical, and the default settings should suffice in almost all circumstances. PATCHTRACK should prove valuable as a an aid in exploration and visualization, particularly in the context of studies involving large-scale movements. The results of the QUILT/PATCHTRACK analysis are consistent across the three proteins studied, suggesting that the trends found are general.

BIBLIOGRAPHY

[1] C. Chothia and J. Janin. Principles of protein-protein recognition. *Nature*, 256:705–708, 1975.

[2] I.M.A. Nooren and J.M. Thornton. Diversity of protein-protein interactions. *EMBO J.*, 22:3486–3492, 2003.

[3] W.E. Meador, A.R. Means, and F.A. Quiocho. Target Enzyme Recognition by Calmodulin: 2.4 Å structure of a Calmodulin complex. *Science*, 257:1251–1253, 1992.

[4] A. Marina, P.M. Alzari, J. Bravo, M. Uriarte, and B. Barcelona. Carbamate kinase: New structural machinery for making carbamoyl phosphate, the common precursor of pyrimidines and arginine. *Prot. Sci.*, 8:934–940, 1999.

[5] J. Heringa and P. Argos. Side-chain clusters in protein structures and their role in protein folding. *J. Mol. Biol.*, 220:151–171, 1991.

[6] L.C. Tisi and P.A. Evans. Conserved Structural Features on Protein Interfaces: Small Exterior Hydrophobic Clusters. *J. Mol. Biol.*, 249:251–258, 1995.

[7] M. Stefani and C.M. Dobson. Protein aggregation and aggregate toxicity: new insights into protein folding, misfolding diseases and biological evolution. *J. Mol. Med.*, 81:678–699, 2003.

[8] J.A. Yankeelov and D.E. Koshland. Evidence for conformation change induced by substrates of phosphoglucomutase. *J. Biol. Chem.*, 240:1593–1602, 1965.

[9] V.N. Uversky. Natively unfolded proteins: a point where biology waits for physics. *Prot. Sci.*, 11:739–756, 2002.

[10] P. Lijnzaad, H.J.C. Berendsen, and P. Argos. A method for detecting hydrophobic patches on protein surfaces. *Proteins*, 26:192–203, 1996.

[11] D. Ferrari, M. Di Valentin, D. Carbonera, A. Merli, Z.-W. Chen, F.S. Mathews, V.L. Davidson, and G.L. Rossi. Electron transfer in crystals of the binary and ternary complexes of methylamine dehydrogenase with amicyanin and cytochrome c551i as detected by EPR spectroscopy. *J. Biol. Inorg. Chem.*, 9:231–237, 2004.

[12] T. Gallagher, P. Alexander, P. Bryan, and G.L. Gilliland. Two crystal structures of the B1 immunoglobulin-binding domain of streptococcal protein G and comparison with NMR. *Biochemistry*, 33:4721–4729, 1994.

[13] R.C.E. Durley, L. Chen, L.W. Lim, F.S. Mathews, and V.L. Davidson. Crystal structure analysis of amicyanin and apoamicyanin from paracoccus denitrificans at 2.0 Angstroms and 1.8 Angstroms resolution. *Prot. Sci.*, 2:739–752, 1993.

[14] B. W. Dijkstra, R. Renetseder, K. H. Kalk, W. G. Hol, and J. Drenth. Structure of porcine pancreatic phospholipase A2 at 2.6 Å resolution and comparison with bovine phospholipase A2. *J. Mol. Biol.*, 168:163–179, 1983.

[15] E. Lindahl, B. Hess, and D. van der Spoel. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Mod.*, 7:306–317, 2001. http://www.gromacs.org/.

[16] H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren, and J. Hermans. *Intermolecular forces*, chapter Interaction models for water in relation to protein hydration, pages 331–342. D. Reidel Publishing company, 1981.

[17] C. Oostenbrink, A. Villa, A.E. Mark, and W.A. van Gunsteren. A biomolecular force field based on the free enthalpy of hydration and solvation: The gromos force-field parameter sets 53A5 and 53A6. *Journal of Computational Chemistry*, 25:1656–1676, 2004.

[18] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M Fraaije. LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry*, 18:1463–1472, 1997.

[19] P. Legendre and L. Legendre. *Numerical Ecology. 2nd English Edition.* Elsevier Science, Amsterdam, 1998.

[20] L. Wall. *http://www.perl.org*.

[21] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005. ISBN 3-900051-07-0.

[22] P. Turner. *Xmgr documentation*. http://plasma-gate.weizmann.ac.il/Xmgr.

[23] R. Sayle and E.J. Milner-White. Rasmol: Biomolecular graphics for all. *Tr. Bioch. Sci.*, 20:374–375, 1995. ftp://ftp.dcs.ed.ac.uk/pub/rasmol/.

[24] P.J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20:53–65, 1987.

[25] J.A. McCammon and S.C. Harvey, editors. *Dynamics of Proteins and Nucleic Acids*. Cambridge University Press, 1987.

[26] G. Hummer, F. Schotte, and P.A. Anfinrud. Unveiling functional protein motions with picosecond X-ray crystallography and molecular dynamics simulations. *Proc. Natl. Acad. Sci.*, 101:15330–15334, 2004.

[27] H. Kovacs, A.E. Mark, and W.F. van Gunsteren. Solvent structure at a hydrophobic protein surface. *Proteins*, 27:395–404, 1997.

[28] A. Luise, M. Falconi, and A. Desideri. Molecular Dynamics Simulation of Solvated Azurin: Correlation between Surface Solvent Accessibility and Water Residence Time. *Proteins*, 39:56–67, 2000.

[29] V.A. Makarov, B.K. Andrews, P.E. Smith, and B.M. Pettitt. Residence Times of Water Molecules in the Hydration Sites of Myoglobin. *Biophys. J.*, 79:2966–2974, 2000.

[30] D. Russo, G. Hura, and T. Head-Gordon. Hydration Dynamics Near a Model Protein Surface. *Biophys. J.*, 86:1852–1862, 2004.

[31] W. Qiu, Y.-T. Kao, L. Zhang, Y. Yang, L. Wang, W.S. Stites, D. Zhong, and A.H. Zewail. Protein surface hydration mapped by site-specific mutations. *Proc. Natl. Acad. Sci.*, 103:13979–13984, 2006.

[32] U. Heugen, G. Schwaab, E. Bründermann, M. Heyden, X. Yu, D. M. Leitner, and M. Havenith. Solute-induced retardation of water dynamics probed directly by terahertz spectroscopy. *Proc. Natl. Acad. Sci.*, 103:12301–12306, 2006.

[33] R.D. Finn, J. Mistry, B. Benjamin Schuster-Böckler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S.R. Eddy, E.L.L. Sonnhammer, and A. Bateman. Pfam: clans, web tools and services. *Nucleic Acids Research*, Database Issue 34:D247–D251, 2006.

[34] B.J. Bahnson. Structure, function and interfacial allosterism in phospholipase A2: insight from the anion-assisted dimer. *Arch. Biochem. Biophys.*, 433:96–106, 2005.

7

DISCUSSION

Knowledge of the structure of proteins is crucial to understanding protein function. Protein folding is an as yet unsolved problem; we cannot predict the fold of a protein sequence having no discernible similarity to any protein of known structure. If the sequence is homologous to other sequences, multiple alignments become better, as does the prediction of secondary structure and solvent accessibility. This in turn improves the structure prediction. The best predictions are obtained when the sequence under study resembles that of an experimentally solved structure. In this case, reasonable models can often be built, and accuracy increases with the the level of homology.

The function of proteins resides largely at their surface. The binding of ligands and of other proteins is determined by the hydrogen-bonding capabilities, shape complementarity, electrostatic potential and hydrophobicity of the interacting partners. These vital details of the protein surface are sensitive to the global fold of the protein, and the accuracy of the surface details therefore depends critically on that of the protein structure. Consequently, experimentally elucidated structures of proteins and protein complexes remain very important for gaining insight into the molecular details of protein function.

Hydrophobicity plays an important role in protein structure. For water soluble proteins, it is the main driving force in folding and protein-protein interaction. The latter phenomenon involves exposure of hydrophobic groups to the solvent, which is energetically unfavourable. The expectation is therefore that proteins will generally avoid exposing local concentrations of hydrophobic surface, as doing so could well lead to local unfolding, or to inadvertent aggregation with other molecules. Such local concentrations of hydrophobic area on the surface of a protein are more commonly called hydrophobic patches, and are frequently seen to have functional significance.

Yet, hitherto, there has not been a systematic description and study of hydrophobic patches on protein surfaces, because no clear-cut definition of "hydrophobic patch" was available. The work described in this thesis provides such a definition of hydrophobic patches and a method, QUILT, to detect them.

A number of other methods have been devised which can function as a background against which to judge the method. Lastly, some parts of the work have engendered controversy, which will be discussed as well. The discussion concludes with suggestions for further work.

SOLVENT ACCESSIBILITY

The foundation of the study of protein surfaces is formed by the calculation of the solvent accessibility. Chapter 2 described the Double Cubic Lattice Method (DCLM) for numerically determining the solvent accessible surface area of proteins. Part or all of the method described therein appear to have been adopted widely.[1–4] A number of approximate methods have been published since,[5–10] yet the DCLM method still finds favour. This must be attributed to the conceptual

simplicity and ease of implementation, but foremost, to speed. A recent survey indicates that, at least for protein structures, it is still the fastest numerical method available.[11]

SURFACE HYDROPHOBICITY

Chapter 3 introduces QUILT, a method to detect hydrophobic patches. The optimal description of surface hydrophobicity depends to some extent on the purpose of the study. Ultimately, a complete understanding of the physics of the protein surface requires a description and parameterization in terms of forces and energies. They could be used in a molecular mechanics setting to quantitatively predict macroscopic observables, particularly the free energy of binding. However, these methods do not lend themselves to visualization. Their main outcome is a single, system-wide quantity such as the free energy, or the dissociation constant. In structural biology such a description is in many cases too coarse, as it is not localized on the protein structure in the way that a patch or a surface is.

OTHER METHODS

There are roughly three approaches to describing protein surface hydrophobicity in a localized fashion: residue counting, lipophilicity potentials, and computational mutation methods.

The simplest approaches resort to residue counting. They are based on the hydrophobicities of amino acids within a certain distance from the residue under consideration.[12–14] They give, at relatively few points of the protein surface, an *ad hoc* measure of hydrophobicity obtained by averaging or summing over relatively few points of the protein surface. As such, these methods are relatively crude.

More detail is afforded by methods that define a *lipophilicity* potential*, in analogy with the electrostatic potential.[15–18] This potential is a so-called *potential of mean force* (PMF), that is, an apparent free energy that reproduces the observed distribution of entities over different phases (in the current case, the distribution of hydrophobic solutes between an apolar environment and water). The form and parameters of these so-called *molecular lipophilicity potentials* (MLP) are empirical. They are usually based on the postulated additivity of atomic hydrophobicity parameters, attenuated by a distance function (usually decaying exponentially, or inversely proportionally). In drug design, the use of potentials of mean force is known as comparative molecular field analysis (CoMFA). These methods enjoy enduring popularity, because their simplicity allows the screening of large numbers of chemical compounds in the hunt for an optimal fit with the active site of the drug target.

*the term lipophilicity is synonymous with hydrophobicity

The MLP potentials are defined as values in space. They have to be contoured or projected before they can be visualized as surfaces. This makes them less suitable as the basis for defining hydrophobic patches, because it requires the setting of a somewhat arbitrary contour level that defines the boundary of a patch.

Lastly, the most advanced treatment of the protein surface is given by the computational mutation methods. They use molecular mechanics simulations to predict the change in free energy of association after computationally mutating exposed amino acid residues into an alanine. By doing a systematic computational mutation scan of the region of interest, the relative binding contribution of each residue is obtained, with near-quantitative accuracy.[19] These scanning methods could be applied to predict the relative contribution of exposed amino acids to the solvation free energy. Compared to the previously described methods, such a treatment of protein surface hydrophobicity could be considered the most physically rigourous. However, the computational mutation methods are normally used to predict affinities in protein-protein and protein-ligand associations, rather than to predict solvation free energies. This again highlights the fact that the description of surface hydrophobicity depends on the application. When the aim is to predict affinities, hydrophobicity alone is not sufficient to capture the intricacies of the molecular interaction.

The objective of a simple concept of "hydrophobicity", for instance like that given by the MLP methods, is prompted by the desire to model hydrophobic interactions with the same rigour and elegance as electrostatic interactions. In the latter case, Coulomb's law provides an exact expression for the electrostatic potential. It is defined as the energy needed to bring one unit of electrical charge from infinity to the point in space under consideration. However, for hydrophobicity, there is no equivalent unit of "hydrophobic charge", nor is there a law governing its behaviour. Although one could postulate a virtual hydrophobic charge (*e.g.*, a methylene group) and a law, it is debatable whether such an approach can accurately capture the details of hydrophobic interactions. For instance, molecular mechanics force fields that omit explicit solvent molecules and instead absorb the apparent attraction between apolar groups into the force field parameters, do not faithfully model the hydrophobic effect. It is more rigourous and accurate to use explicit water to achieve proper modelling of hydrophobic phenomena and other details of the molecular system.

Based on the these considerations, any measure of hydrophobicity, and any definition of hydrophobic patches based on it, will necessarily contain an element of subjectivity. The premise of the QUILT patch detection method is that surface hydrophobicity is best described as contiguous areas of carbon and sulfur atoms. This atomistic approach is a simplification from a physico-chemical point of view, as hydrophobicity arises from the combined action of apolar groups, polar groups, surface geometry, and, most of all, water behaviour. A more detailed description could, for instance, also include partial charges and polarization. This would allow for the occasionally observed hydrogen bonds with the conjugated systems of ty-

rosine, phenylalanine and tryptophan. This would however affect the simplicity of the current approach.

With regard to the above classification of surface hydrophobicity descriptions, QUILT takes a middle ground. It is much more detailed than the residue counting methods, yet not as detailed as the free energy methods. Its additional strengths are that it is robust, has no parameters*, is fast, conceptually simple, and especially well suited to visualization purposes.

THE METHOD BY EISENHABER AND ARGOS

A patch detection method nearly identical to that described in Chapter 3 was implemented by Eisenhaber and Argos.[20] They use an analytical surface calculation which, unlike QUILT, does not recover hydrophobic surface area that gets buried as a result of the polar expansion. In a sense, there is less need for that, because they find a much smaller polar expansion radius of 0.4 Å to be optimal. The reasoning is as follows. Extending, by this value, all the polar atoms in four selected protein structures yields the same loss in hydrophobic surface area as that seen in a model of the hydrated proteins. This model is obtained by placing waters of radius 1.9 Å at positions predicted to be best suited to fixed hydration waters. Hence, specific hydration of some of the polar atoms is mimicked by expanding all polar atoms, thus taking into account the averaging caused by the motion of the protein.

While this approach is more sophisticated than simply adding the water radius to polar atoms as advocated in Chapter 3, it too is speculative: Four protein structures may not be representative, 1.9 Å as the radius of the hydration waters is unusual (1.4 Å is nearly universally used), and it is debatable whether hydration waters can be considered a separate species.

In a sense, the value of the expansion radius determines the resolution at which surface hydrophobicity is described. Using a small expansion radius yields low-resolution patches. They have on average a greater area than those obtained with a large expansion radius, but they can contain many more islands of polar atoms. In contrast, using a large expansion radius results in smaller, but more concise patches that rarely contain polar atoms. As argued in Chapter 3, the value of the polar expansion radius cannot be derived from first principles. QUILT's default value of 1.4 Å is suggested by hydration, and produces patches that correspond well to those selected visually by a human observer. A maximum in the ratio of hydrophobic of hydrophilic patches occurs roughly around 1.4 Å (Chapter 3, Fig. 4), stressing the optimality for purposes of visualization.

The distribution of patch sizes is roughly exponential (more on this below). Interestingly, the "decay constant" is not affected much by the choice of the expansion radius, as shown in Fig. 1. The polar expansion radius uniformly affects small and large patches.

*excepting the polar expansion radius, which by default is 1.4 Å.

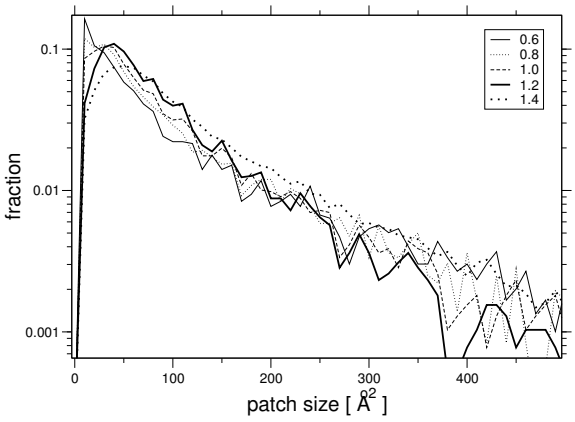

Figure 1: *The distribution of patch sizes, as influenced by the choice of polar expansion radius (indicated in legend). The figure is based on a set of 78 monomeric proteins (a subset of those used in Chapter 4).*

PATCH SIZE DISTRIBUTIONS

In Chapter 4, the QUILT method was applied to a large number of monomeric proteins. It showed that small patches are abundant, medium patches less so, and large patches not at all. In fact, the patch sizes are exponentially distributed, at least over a large part of their range. This might be interpreted as a Boltzmann distribution. In a Boltzmann equilibrium, the ratio between entities in different states is described by the equation:

$$\frac{n_2}{n_1} = \mathrm{e}^{-\frac{E_2 - E_1}{kT}}$$

where n_1 and E_1 are the number of entities in state 1, and the energy of that state, and likewise for n_2 and E_2; k is the Boltzmann constant, and T the temperature. This is a fundamental result from statistical mechanics, and gouverns for instance the atmospheric distribution of gases and the concentration of chemical reactants that are in equilibrium. It can be summarized by stating that large energy differences yield overpopulated lower energy states, but raising the temperature lessens this bias.

The Boltzmann formula also forms the basis for the derivation of the *knowledge based potentials* (see Chapter 1) and potentials of mean force. The reasoning is that the observed distribution of entities is caused by the action of an apparent energy (or its derivative, force) acting on the ensemble of entities. By fitting energies to the logarithm of the abundances, one can infer the underlying energy parameters. For instance, Casari and Sippl[21] deduce an apparent distance-dependent

force between the 20 amino acid types, based on a sample of 88 protein structures in the Protein Database. Using this potential, they can distinguish native from misfolded protein structures.

In the case of the QUILT-patches, the matter is much more straightforward, because the energetic cost of hydration is proportional to their area, as

$$\Delta G_{solv} = \sigma_{solv} \cdot A$$

with ΔG_{solv} the energy of transfer of hydrophobic solvation per mol, and σ_{solv} the specific hydrophobic surface solvation free energy in calories per mol per Å^2. So if patch sizes are distributed as

$$f = \mathrm{e}^{-\frac{\Delta G}{RT}} = \mathrm{e}^{-\frac{A \cdot \sigma_{solv}}{RT}}$$

(switching to molar quantities using the gas constant, R), then the logarithm of the histogram would behave as

$$\log f \sim -A \cdot \frac{\sigma_{solv}}{RT}$$

hence

$$\sigma_{solv} = -slope(\log(f)) \cdot RT$$

The temperature would be the ambient temperature of the organism producing the proteins, *i.e.* 300 K for mesophilic species.

If I apply, by way of example, the above formula to the distribution shown before in Fig. 2, I obtain a σ_{solv} of around 6 cal/Å^2 mol. This is suspiciously close to experimental values, which range from 8 to 73 cal/Å^2 mol; see Ref. 22. The big question now is: is it appropriate to interpret the distribution of patch sizes in this way? Eisenhaber answers it in the affirmative; he arrives at a σ_{solv} value of around 18 cal/Å^2 mol using this approach.[20] Differences between our respective σ_{solv} values are probably due to using different data sets and methods, but what matters is that the value is so close to experimental values. Is this proof of the validity of the Boltzmann interpretation?

I do not think so, for the following reasons. Firstly, each separate protein has a roughly exponential distribution of patch sizes (data not shown), and it is difficult to see how evolution would produce, in each protein structure, such a distribution of hydrophobicity over the protein surface. It would imply that on an evolutionary time scale, patches on the surface of one protein are in a dynamical equilibrium of expanding and shrinking. Secondly, the surface of proteins from thermophilic organisms is generally slightly more hydrophilic than that of mesophilic species.[23, 24] This is at odds with the Boltzmann interpretation, because at higher temperatures,

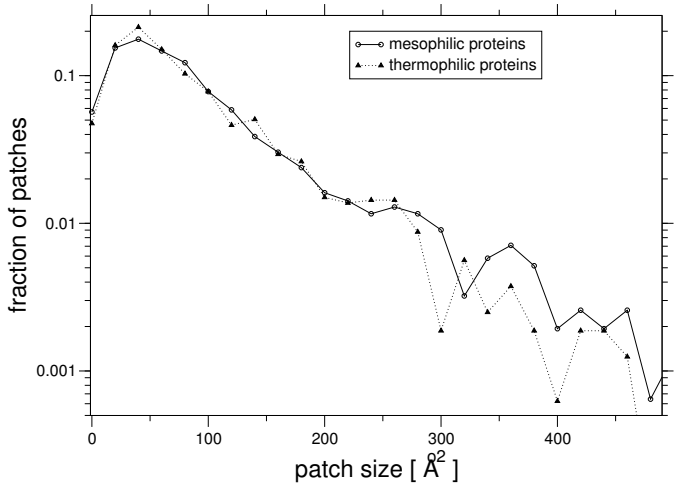

Figure 2: *The distribution of patch sizes for mesophilic and thermophilic proteins.*

the energetic cost of exposing apolar surface to the solvent can be more easily met. This would imply generally larger patches, whereas in fact slightly smaller ones are observed. This can be seen in Fig. 2, which shows the size distribution of patches on structures of a number of paired orthologues from mesophilic and thermophilic organisms.[24] The thermophilic data decays slightly more steeply than that of the mesophilic organisms as a result of the slightly higher fraction of polar groups. Conversely, if we base the estimation of σ_{solv} on the thermophilic structures, using $T = 343\ K$ as the lower limit of the thermophile's temperature range, a σ_{solv} of around 8 cal/Å^2 mol is obtained. This is inconsistent with the previous outcome of around 6 cal/Å^2 mol. Although this is a preliminary analysis with little data, it would seem to contradict the Boltzmann interpretation. Lastly, if the protein surface is randomized prior to patch detection, a nearly identical graph results (data not shown). This indicates that the characteristics of the patch size distributions are dictated exclusively by the fraction of polar atoms on the surface, rather than by the details of the protein sequence. In other words, not the patches, but the overall hydrophobicity of a protein's surface matters.

Many other protein structure statistics have been interpreted as reflecting or representing Boltzmann equilibria; examples are given in Ref. 25. Frequently, this approach enables the recognition of incorrect predictions of the protein fold, or details thereof.[21, 25, 26] However, the energies obtained in such studies are fictitious; they do not correspond to macroscopic energies. It is somewhat puzzling that the σ_{solv} obtained from the distribution of patch sizes is in the same order of

magnitude as the experimental solvation free energies, and while this deserves an explanation, there is currently not enough evidence to interpret the patch size distribution as reflecting an evolutionary Boltzmann equilibrium. More work will be required to settle this question. Energetic considerations will be an essential part of the explanation, as highlighted by the fact that the area of the largest patch found on a protein is not dependent on the size of the protein (see Chapter 4). The avoidance of local concentrations of hydrophobic surface area is likely to be an adaptation that prevents local unfolding and aggregation caused by the energetics of hydrophobic exposure.

INTERFACES AND PATCH COMPLEMENTARITY

The study of interfaces in protein complexes is unabatedly popular.[27–37] Surveying it has become more specialized in recent years, and has diverged into that of interfaces in transient complexes,[33, 38, 39] homodimers,[28, 31, 37] as well as domain-domain interfaces[40, 41]*. Various aspects of our understanding of protein association have been deepened, enabling, amongst others, improved recognition of protein interfaces. For instance, the Abagyan group recently demonstrated a 97% success rate in predicting the protein interface.[36] Another important realization is that protein complexes are not optimized for stability.[29]

In Chapter 5, QUILT was applied to the interfaces of obligate complexes. One of the results was that hydrophobic patches are not in contact across the interface. While perhaps a surprising result, I argued that this is to be expected. Hydrophobicity is not a simple attraction between apolar groups, but rather the result of water's tendency not to sacrifice hydrogen bonds nor entropy by shying away from of such groups. However, in the interface between complexed subunits, water is is already nearly completely absent, and the interface looks much like the protein interior. The cost of exposing hydrophobic surface area to the solvent is avoided by the subunits mutually burying it as a result of the complexation, but there is no requirement that the hydrophobic patches pair across the interface.

However, all studies looking at complementarity of hydrophobicity across interfaces do report the preference for hydrophobic moieties to match up.[12, 13, 37, 39, 42] In all cases, this appears to be due to the approach taken, which is based on a relatively course-grained description of the interface based on complete residues. The description afforded by the QUILT-patches is at a finer level, and shows no such complementarity.

The situation can perhaps be compared to that of the electrostatic similarity across subunit interfaces. Here, there is little complementarity at the level of individual charges, but when described in terms of electrostatic potential, there is good correspondence.[43] In this case, it is clear that the proper physical description is that of the full electrostatic potential. However, in the case of hydrophobic patches on subunit interfaces, it is less clear what the proper physical description

*Domain-domain interfaces are those between domains within one polypeptide chain.

is, so further research is required to decide this matter. One would, however, have expected the above studies to at least comment on the fact that hydrophobic complementarity is found, yet not strictly required for the stability of the association. For instance, a possible explanation could be that such complementarity is needed for steering during the process of association, or that it is simply a by-product of the tendency to form hydrogen bonds across the interface.

DYNAMICS OF PATCHES

Protein structures are inherently dynamic, and in many cases, such dynamism is functionally significant. It is therefore interesting to study the dynamics of the hydrophobic patches. This is the subject of Chapter 6, in which QUILT is used to study patches on "snapshots" of protein structures taken from Molecular Dynamics simulations. It requires an additional procedure called PATCHTRACK which produces *patch runs*, series of QUILT-patches that have been matched across the consecutive snapshots of a molecular dynamics trajectory. If PATCHTRACK cannot find, in the next snapshot, a suitable patch to prolong a patch run, the patch run ends. The results reveal that the QUILT-patches are subject to rather large fluctuations (around 25% of their size), and are, as a result, rather unstable: the average live-span of patch runs is only around 4 ps. The volatility arises as a result of the fluctuations that are generally caused by the addition (or removal) of complete atoms or even complete side chains to (or from) an ongoing patch run. This is typically due to relatively small changes in the positions of the apolar atoms used by QUILT to delineate the hydrophobic patches.

The instability of the patches might therefore seem an artifact of QUILT. However, any definition of hydrophobic patches based on contiguous surfaces will involve discontinuities in a dynamical context, and will as a result be inherently noisy. This noise can be reduced by taking (many) more snapshots of the structure during the simulations. This should allow the matching procedure to achieve longer patch runs. However, the cost of this is considerable, and the discontinuities do not disappear. The pragmatic stance is to sort results, and discard those deemed too noisy. This approach is taken by clustering the patch runs into so-called recurrent patches. In most cases, only the most persistent recurrent patches (those represented by a "live patch run" during most of the time) will be of interest.

However, even for the highly live recurrent patches, noise is inevitable. Frequently, the patch runs constituting one recurrent patch often show brief temporal overlaps, particularly at their starts and ends. These temporal overlaps are caused by the fortuitous split of a smaller "satellite" off the main patch. The main patch and the satellite patch run in parallel for a few picoseconds, the main patch comes to an end, and the satellite patch tends to get most of the atoms that used to be belong to the main patch. These fortuitous overlaps are probably difficult to mend in a general way. However, the problem does not appear to be serious, as it concerns only 3% of the simulation time.

The extent to which patch dynamics are important for function is difficult to assess. The largest patch in amicyanin is very persistent, and is known to be involved in the binding of methylamine dehydrogenase. This case of biologically relevant patch persistence does at least suggest an avenue of further investigation.

FUTURE WORK

The current work offers a number of opportunities for further study. As discussed in one of the preceding sections, QUILT has proven useful for visualization and exploration, but it would certainly be interesting to investigate the connections with the more physico-chemical aspects of hydrophobicity, particularly those given by the MLP methods. In all likelihood, QUILT-patches will correlate strongly with those found by other methods, while at the same time having the benefit of conceptual simplicity and speed. However, establishing this is something that would require a careful comparison. Such an investigation would probably also be the right opportunity to resolve the noted discrepancy, between QUILT and other methods, regarding hydrophobic complementarity in interfaces of protein complexes.

The study of patches on subunit interfaces was limited to obligate complexes. It is known that the more transient non-obligate complexes are hydrophobic, but to a lesser degree.[32, 33, 44] It will be attractive to know exactly how hydrophobic patches behave in these cases. This could be even more interesting than patches on interfaces of obligate complexes, because those of transient complexes are more difficult to recognize and to dock. And, whereas there is no patch complementarity for obligate complexes (see Chapter 5), there may be for transient complexes. If so, this should be helpful in molecular docking endeavours.

Thornton and co-workers have also looked into the details of protein domain interfaces.[40] This part of the protein surface (if "surface" is an appropriate term) has so far received scant attention, and could also be subjected to QUILT. The expectation is that results will be quite similar to those of the interface patches in obligate complexes.

A major topic not addressed in this thesis is the evolutionary biology of the QUILT-patches. This will first of all require that patches be redefined in terms of amino acid residues, because in the current definition, only atoms are considered. Having cast the QUILT-patches in terms of residues, a number of questions can be posed. To what extent are the patches conserved? Is there a relation between their size and their degree of conservation? Is the significance level given by the bootstrap method described in Chapter 3 a good indicator of functional significance? Can any patch-forming tendency be distinguished in sequences alone? The conservation of the majority of hydrophobic residues is usually due to them forming part of the hydrophobic core, so chances of disentangling patch participation from core membership must be deemed minute. Note that even interfaces cannot be predicted from sequence with any accuracy.[45]

Another venue of inquiry is suggested by the apparent Boltzmann distribution

of patch sizes, as discussed above. Is it justified to interpret the observed distribution as such? This will require consideration of protein structures from organisms living at a wide range of ambient temperatures. As discussed above, the slope of the patch size distribution depends directly on the fraction hydrophobic surface area per protein. Therefore, not the shape of the patch size distribution, but rather the fraction hydrophobic surface area is the fundamental quantity to be studied. It should be revealing to investigate its distribution across the universe of protein structures, and to see whether it can be interpreted as a Boltzmann distribution.

The study of the dynamics of the hydrophobic patches using Molecular Dynamics could be extended to include much longer simulations, and many more structures, especially those with functionally relevant hydrophobic patches. I expect these to be more persistent than fortuitous patches. Molecular Dynamics simulations of protein-ligand or protein-protein association is another area in which QUILT holds great potential; it should help establish the functional role of patch dynamics. Simulating homologous structures will enable the study of the conservation of the patch dynamics, and also could shed light on their significance.

Lastly, the current work, like many other theoretical studies, suffers the drawback of not having been confirmed by experiment. One of our observations, for instance, is that there is little patch complementarity across subunit interfaces, nor should there be. It may be possible to test this claim by modifying the interface to alter the complementarity, and experimentally determining the change in dissociation constant. This protein engineering will have to be preceded by a very careful molecular modelling of the likely result, because disruption of the shape complementarity, or of the hydrogen-bonding,[27, 46, 47] will easily swamp any predicted lack of effect.

Another issue that merits experimental attention is that of protein aggregation. Mutating a hydrophilic surface residue into a hydrophobic one generally results in little destabilization. However, if a mutation increases the area of an already large hydrophobic patch, the likely consequence would be a greater tendency to aggregate as compared to a similar mutation enlarging a smaller hydrophobic patch on the same protein. Conversely, introducing a hydrophilic residue in a large hydrophobic patch should reduce aggregation tendency. These predictions can be verified relatively easily.

To conclude the suggestions for verification, in Chapter 6 the prediction is made that the Leu58,Phe94 patch in phospholipase A_2 is functionally important, based on patch size, persistence and conservation. If borne out by experiment, the tools developed in this thesis would have found something that appears to have been overlooked in over 30 years of study into this intriguing protein.

In conclusion, the current work describes a method to study an aspect of protein structures that was hitherto not well described, namely that of hydrophobic surface patches. The method has proven useful in the research of a wide variety of individ-

ual protein structures.[38, 48–55] Its application in surveys of protein structures has provided insights regarding surface hydrophobicity in general. These observations have been compared and used in other work[6, 18, 34, 39, 45, 56–60] The method and the surveying results are a valuable tool in the study of protein-protein interactions, protein stability and molecular simulation.

BIBLIOGRAPHY

[1] R. Abagyan. Personal communication, 1996.

[2] E. Lindahl, B. Hess, and D. van der Spoel. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Mod.*, 7:306–317, 2001. http://www.gromacs.org/.

[3] E.J. Sorin, Y.M. Rhee, M.R. Shirts, and V.S. Pande. The Solvation Interface is a Determining Factor in Peptide Conformational Preferences. *J. Mol. Biol.*, 356:248–256, 2006.

[4] F. Eisenmenger, U.H.E. Hansmann, S. Hayryan, and C.-K. Hu. An enhanced version of smmp – open-source software package for simulation of proteins. *Comp. Phys. Comm.*, 174:422–429, 2006.

[5] M.F. Sanner, A.J. Olson, and J.C. Spehner. Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers*, 38:305–320, 1996.

[6] R.R. Gabdoulline and R.C. Wade. Analytically defined surfaces to analyze molecular interaction properties. *J. Mol. Graph.*, 14:341–353, 1996.

[7] Y. N. Vorobjev and J. Hermans. SIMS: computation of a smooth invariant molecular surface. *Biophys. J.*, 73:722–732, 1997.

[8] J. Weiser, P.S. Shenkin, and W.C. Still. Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO). *Journal of Computational Chemistry*, 20:217–230, 1999.

[9] J. Weiser, P.S. Shenkin, and W.C. Still. Fast, approximate algorithm for detection of solvent-inaccessible atoms. *Journal of Computational Chemistry*, 20:586–596, 1999.

[10] S. Bhat and E.O. Purisima. Molecular surface generation using a variable-radius solvent probe. *Proteins*, 62:244–261, 2005.

[11] P.S. Ivanov, P. Jain, D. Osei-Kuffuor, V.S. Stanev, R. Tripathi, P.N. Zhivkov, and H.J. Bernstein. Comparison of recent algorithms for computing molecular surfaces. In *Proc. Natl. Conf. Undergrad. Res.* National Conferences on Undergraduate Research, 2003.

[12] A.P. Korn and R.M. Burnett. Distribution and complementarity of hydropathy in multi-subunit proteins. *Proteins*, 9:37–55, 1991.

[13] L. Young, R.L. Jernigan, and D.G. Covell. A role for surface hydrophobicity in protein-protein recognition. *Prot. Sci.*, 3:717–729, 1994.

[14] S. Jones and J.M. Thornton. Analysis of Protein-Protein Interaction Sites using Surface patches. *J. Mol. Biol.*, 272:121–132, 1997.

[15] P. Furet, A. Sele, and N.C. Cohen. 3D molecular lipophilicity potential profiles: a new tool in molecular modeling. *J. Mol. Graph.*, 6:182–189, 1988.

[16] G.E. Kellogg, S.F. Semus, and D.J. Abraham. HINT: A new method of empirical hydrophobic field calculation for CoMFA. *J. Comput.-Aided Mol. Des.*, 5:545–552, 1991.

[17] R. Jäger, F. Schmidt, B. Schilling, and J. Brickmann. Localization and quantification of hydrophobicity: The molecular free energy density (MolFESD) concept and its application to sweetness recognition. *J. Comput.-Aided Mol. Des.*, 14:631–646, 2000.

[18] J. Fernandez-Recio, M. Totrov, C. Skorodumov, and R. Abagyan. Optimal docking area: A new method for predicting protein-protein interaction sites. *Proteins*, 58:134–143, 2005.

[19] W. Wang, O. Donini, C. M. Reyes, and P. A. Kollman. Biomolecular simulations: recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions. *Ann. Rev. Biophys. Biomol. Struct.*, 30:211–243, 2001.

[20] F. Eisenhaber and P. Argos. Hydrophobic regions on protein surfaces: definitions based on hydration shell structure and a quick method for their computation. *Prot. Engng.*, 9:1121–1133, 1996.

[21] G. Casari and M.J. Sippl. Structure-derived hydrophobic potential. Hydrophobic potential derived from X-ray structures of globular proteins is able to identify native folds. *J. Mol. Biol.*, 224:725–732, 1992.

[22] A.H. Juffer, F. Eisenhaber, S. J. Hubbard, D. Walther, and P. Argos. Comparison of atomic solvation parametric sets: Applicability and limitations in protein folding and binding. *Prot. Sci.*, 4:2499–2509, 1995.

[23] G. Vogt, S. Woell, and P. Argos. Protein thermal stability, hydrogen bonds, and ion pairs. *J. Mol. Biol.*, 269:631–643, 1997.

[24] S. Kumar, C.-J. Tsai, and R. Nussinov. Factors enhancing protein thermostability. *Prot. Engng.*, 13:179–191, 2000.

[25] D. Shortle. Propensities, probabilities, and the Boltzmann hypothesis. *Prot. Sci.*, 12:1298–1302, 2003.

[26] H. Lu and J. Skolnick. A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins*, 44:223 – 232, 2001.

[27] L. Lo Conte, Chothia C., and J. Janin. The atomic structure of protein-protein recognition sites. *J. Mol. Biol.*, 285:2177–2198, 1999.

[28] W.S.J. Valdar and J.M. Thornton. Protein-protein interfaces: Analysis of amino acid conservation in homodimers. *Proteins*, 42:108–124, 2000.

[29] N. Brooijmans, K.A. Sharp, and I.D. Kuntz. Stability of macromolecular complexes. *Proteins*, 48:645–653, 2002.

[30] P. Chakrabarti and J. Janin. Dissecting protein-protein recognition sites. *Proteins*, 47:334–343, 2002.

[31] R.P. Bahadur, P. Chakrabarti, F. Rodier, and J. Janin. Dissecting subunit interfaces in homodimeric proteins. *Proteins*, 53:708–719, 2003.

[32] I.M.A. Nooren and J.M. Thornton. Diversity of protein-protein interactions. *EMBO J.*, 22:3486–3492, 2003.

[33] I.M.A. Nooren and J.M. Thornton. Structural Characterisation and Functional Significance of Transient Protein-Protein Interactions. *J. Mol. Biol.*, 325:991–1018, 2003.

[34] M. Guharoy and P. Chakrabarti. Conservation and relative importance of residues across protein-protein interfaces. *PNAS*, 102:15447–15452, 2005.

[35] H. Ponstingl, T. Kabir, D. Gorse, and J.M. Thornton. Morphological aspects of oligomeric protein structures. *Prog. Biophys. Mol. Biol.*, 89:9–35, 2005.

[36] A.J. Bordner and R. Abagyan. Statistical analysis and prediction of protein-protein interfaces. *Proteins*, 60:353–366, 2005.

[37] Y. Tsuchiya, K. Kinoshita, and H. Nakamura. Analyses of homo-oligomer interfaces of proteins from the complementarity of molecular surface, electrostatic potential and hydrophobicity. *Prot. Engng Des. Sel.*, 19:421–429, 2006.

[38] H. Neuvirth, R. Raz, and G. Schreiber. ProMate: A Structure Based Prediction Program to Identify the Location of Protein-Protein Binding Sites. *J. Mol. Biol.*, 338:181–199, 2004.

[39] S. Ansari and V Helms. Statistical analysis of predominantly transient protein-protein interfaces. *Proteins*, 61:344–355, 2005.

[40] S. Jones, A. Marin, and J.M. Thornton. Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. *Prot. Engng.*, 13:77–82, 2000.

[41] W.K. Kim and J.C. Ison. Survey of the geometric association of domain-domain interfaces. *Proteins*, 61:1075–88, 2005.

[42] F. Glaser, D.M. Steinberg, I.A. Vakser, and N. Ben-Tal. Residue Frequencies and Pairing Preferences at Protein-Protein Interfaces. *Proteins*, 43:89–102, 2001.

[43] A.J. McCoy, V. Chandana Epa, and P.M. Colman. Electrostatic Complementarity at Protein/Protein Interfaces. *J. Mol. Biol.*, 268:570–584, 1997.

[44] S. De, O. Krishnadev, N. Srinivasan, and N. Rekha. Interaction preferences across protein-protein interfaces of obligatory and non-obligatory components are different. *BMC Struct. Biol.*, 5:15–31, 2005.

[45] D.R. Caffrey, S. Somaroo, J.D. Hughes, J. Mintseris, and E.S. Huang. Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Prot. Sci.*, 13:190–202, 2004.

[46] J. Janin, S. Miller, and C. Chothia. Surface, subunit interfaces and interior of oligomeric proteins. *J. Mol. Biol.*, 204:155–164, 1988.

[47] J. Jones and J.M. Thornton. Protein-protein interactions: a review of protein dimer structures. *Prog. Biophys. Mol. Biol.*, 63, 1995.

[48] L.D. Creveld, A. Amadei, R.C. van Schaik, H.A.M. Pepermans, J. de Vlieg, and H.J.C. Berendsen. Identification of functional and unfolding motions of cutinase as obtained from molecular dynamics computer simulations. *Proteins*, 33:253–264, 1998.

[49] M.J. Pandya, P.B. Sessions, R.B. Williams, C.E. Dempsey, A.S. Tatham, P.R. Shewry, and A.R. Clarke. Structural characterization of a methionine-rich, emulsifying protein from sunflower seed. *Proteins*, 38:341–349, 2000.

[50] J. Piehler, L.C. Roisman, and G. Schreiber. New Structural and Functional Aspects of the Type I Interferon-Receptor Interaction Revealed by Comprehensive Mutational Analysis of the Binding Interface. *J. Biol. Chem.*, 275:40425–40433, 2000.

[51] T. Granier, B. Gallois, B. Langlois d'Estaintot, A. Dautant, J.-M. Chevalier, J.-M. Mellado, C. Beaumont, P. Santambrogio, P. Arosio, and G. Precigoux. Structure of mouse L-chain ferritin at 1.6 Å resolution. *Acta Cryst.*, D57:1491–1497, 2001.

[52] M.M. Mullen, K.M. Haan, R. Longnecker, and T.S. Jardetzky. Structure of the Epstein-Barr Virus gp42 Protein Bound to the MHC Class II Receptor HLA-DR1. *Molecular Cell*, 9:375–385, 2002.

[53] N. Pokala and T.M. Handel. Energy Functions for Protein Design: Adjustment with Protein-Protein Complex Affinities, Models for the Unfolded State, and Negative Design of Solubility and Specificity. *J. Mol. Biol.*, 347:203–227, 2005.

[54] T.J. Brett, V. Legendre-Guillemin, P.S. McPherson, and D.H. Fremont. Structural definition of the F-actin-binding THATCH domain from HIP1R. *Nature Struct. Mol. Biol.*, 13:121–130, 2006.

[55] A. Kriško and C. Etchebest. Theoretical model of human apolipoprotein B100 tertiary structure. *Proteins*, 66:0–0, 2007.

[56] Y.-K. Cheng and P.J. Rossky. The effect of vicinal polar and charged groups on hydrophobic hydration. *Biopolymers*, 50:742–750, 1999.

[57] G. D'Alessio. The evolutionary transition from monomeric to oligomeric proteins: tools, the environment, hypotheses. *Prog. Biophys. Mol. Biol.*, 72:271–298, 1999.

[58] I. Angrand, L. Serrano, and E. Lacroix. Computer-assisted re-design of spectrin SH3 residue clusters. *Biomol. Engng.*, 18:125–134, 2001.

[59] M. Nayal and B. Honig. On the nature of cavities on protein surfaces: Application to the identification of drug-binding sites. *Proteins*, 63:892–906, 2006.

[60] J. Hoskins, S. Lovell, and T.L. Blundell. An algorithm for predicting protein-protein interaction sites: Abnormally exposed amino acid residues and secondary structure elements. *Prot. Sci.*, 15:1017–1029, 2006.

SUMMARY

Hydrophobicity is a prime determinant of the structure and function of proteins. It is the driving force behind the folding of soluble proteins, and when exposed on the surface, it is frequently involved in recognition and binding of ligands and other proteins. The energetic cost of exposing hydrophobic surface is proportional to its area, and the question arises to what extent proteins can tolerate large hydrophobic patches on their surfaces. The current thesis is a study into such patches.

Chapter 1 provides a general introduction into protein surface hydrophobicity.

Chapter 2 describes a numerical algorithm for calculating the solvent accessible surface area. It samples the protein surface in Shrake & Rupley fashion: representing atoms as spherical distributions of points and summing the points that are not buried by any atoms. A number of optimization strategies is applied, yielding an exceptionally fast method. The quality of spherical point distributions is assessed, and a novel, optimal tessellation of the unit sphere is found.

The accessible surface calculation method developed in Chapter 2 forms the basis of the hydrophobic patch detection algorithm called QUILT, presented in Chapter 3. The assumption is that hydrophobic surface area is synonymous with solvent accessible carbon and sulfur atoms. Connecting contiguous apolar atoms is not enough to delineate hydrophobic patches, because the relatively strong hydrophobicity of the protein surface (around 60%) results in one large hydrophobic surface. This surface spans the entire protein, and is dotted with polar "islands" formed by the hydrophilic atoms, with hydrophobic connections through variously sized "channels" between these islands. To delineate the hydrophobic patches, the channels are closed off by temporarily expanding the solvent-accessible polar atoms. This way, the hydrophobic surface neatly divides into proper patches which are subsequently identified, and adjacent surface area lost due to the polar expansion is added back to the patches thus obtained. Only the largest patches, having sizes exceeding expectation (based on randomizing the protein's surface), are deemed meaningful. The method is applied to a small number of structures to demonstrate the validity and utility of the method.

In Chapter 4, the QUILT method is applied to a large sample of monomeric proteins, in order to survey general trends in the distribution of patch sizes on proteins. The largest patch on each individual protein averages around 400 $Å^2$, but can range from 200 to 1200 $Å^2$. Interestingly, these areas do not correlate with the sizes of the proteins, and only weakly with their apolar surface fraction.

Trends regarding patch size distribution, amino acid composition and preference, sequential vicinity, secondary structure and mobility are discussed as well.

Chapter 5 is devoted to a survey similar to that described in Chapter 4, but here, the interfaces of obligate oligomeric proteins are studied. As before, trends regarding amino acid composition and preference and patch size distribution are described. The largest or second largest patch on the accessible surface of the entire subunit was involved in multimeric interfaces in 90% of the cases, in agreement with interfaces being generally more hydrophobic than the rest of the protein surface. However, hydrophobic patches are not complementary: they are not preferentially in contact across associating subunits. This is perhaps surprising, but is to be expected, because the free energy of subunit association, as far as the hydrophobic patches are concerned, is largely due to the shielding of apolar area from the solvent, rather than from gaining hydrophobic contacts.

To gain insight into the dynamic behaviour of hydrophobic patches, QUILT is applied to molecular dynamics simulations of three different protein structures. This is the subject of Chapter 6. The analysis requires an additional method to relate QUILT-patches across time frames of the trajectory, which is described as well. The resulting *patch runs* show that the area fluctuations are considerable, at around 25% of their size. The most frequently occurring mean patch size is approximately 50 $Å^2$, but can reach around 400 $Å^2$. An uninterrupted patch run can last up to 150 picoseconds, but, owing to protein mobility, is generally much shorter at around 4 ps. There is no clear relation between patch run durations and their average size, but long-lasting patch runs have smaller fluctuations. Although the formalism would allow this, the patches do not “wander” over the protein surface, indicating that they are genuine surface features. When the patch runs are clustered, the truly persistent patches called recurrent patches are obtained. Only about 25% of them have a strong ”liveness”, that is, are represented by an actual patch run most of the time. In amicyanin, the method detects the hydrophobic patch known to be involved in the binding of methylamine dehydrogenase. In phospholipase A_2, a large persistent patch consisting of Leu58 and Phe94 is found, the likely functional relevance of which appears to be novel.

Chapter 7 concludes with a discussion of the merits of QUILT, its relationship with other methods, and results obtained by other researchers. Lastly, suggestions for further research are presented.

Hydrofobiciteit ("waterafstotendheid") speelt een belangrijke rol in de structuur en functie van eiwitten. Het is de drijvende kracht achter de vouwing van wateroplosbare eiwitten, en hydrofobiciteit aan het oppervlak is veelal betrokken bij het herkennen en binden van liganden en andere eiwitten. De energetische kosten van het blootstellen van hydrofoob oppervlak aan water zijn evenredig met de omvang, en de vraag is in welke mate eiwitten grote hydrofobe gebieden op hun oppervlak kunnen verdragen.

Hoofdstuk 1 geeft een algemene inleiding in de rol van hydrofobiciteit aan het eiwitoppervlak.

Hoofdstuk 2 beschrijft een numeriek algoritme voor de berekening van het watertoegankelijke eiwitoppervlak. Het bemonstert het eiwitoppervlak met de Shrake & Rupley methode: atomen worden gerepresenteerd door bolvormige verdelingen van punten, en de punten die niet begraven liggen in andere atomen worden opgeteld. Een aantal optimaliseringsstrategieën wordt toegepast, waardoor een uitzonderlijk snelle methode is verkregen.

De in Hoofdstuk 2 ontwikkelde methode voor de berekening van het toegankelijk oppervlak vormt de grondslag van QUILT. Dit is een methode voor het opsporen van hydrofobe gebieden op het eiwitoppervlak (hierna *patches* genoemd), en wordt gepresenteerd in Hoofdstuk 3. De aanname is dat hydrofoob oppervlak hetzelfde is als watertoegankelijke koolstof- en zwavelatomen. Het met elkaar verbinden van naast elkaar gelegen apolaire atomen is echter niet voldoende om de patches af te bakenen, omdat de verhoudingsgewijs hydrofobe aard van het eiwitoppervlak (ongeveer 60%) dan uitmondt in één groot hydrofoob gebied. Dit gebied omvat het gehele eiwitmolecuul, waarin polaire "eilandjes" liggen die gevormd worden door de hydrofiele atomen, terwijl er hydrofobe "kanalen" van verschillende omvang tussen de eilandjes door lopen. Om nu de patches af te bakenen worden deze kanalen tijdelijk afgesloten door de watertoegankelijke polaire atomen te vergroten. Op deze manier raakt het hydrofobe oppervlak mooi verdeeld in afzonderlijke gebieden, die vervolgens worden opgespoord. Aangrenzend hydrofoob oppervlak dat door de polaire vergroting verloren was gegaan wordt weer aan de verkregen patches toegevoegd. Slechts de grootste patches, met een omvang groter dan de verwachtingswaarde (verkregen door het *randomizeren* van het eiwitoppervlak) worden van betekenis geacht.

In Hoofdstuk 4 wordt de QUILT methode toegepast op een groot aantal mo-

nomere eiwitten, om zicht te krijgen op de verdeling van patchgrootten bij eiwitten in het algemeen. De grootste patch van een eiwit is gemiddeld ongeveer 400 $Å^2$ groot, maar kan van 200 to 1200 $Å^2$ groot zijn. Het is opmerkelijk dat deze oppervlakten niet correleren met de eiwitgrootten en slechts zwak met hun percentage hydrofoob oppervlak. Tevens worden trends betreffende de grootteverdeling van patches, hun aminozuursamenstelling en -voorkeur, sequentieafstand, secundaire structuur en beweeglijkheid besproken.

Hoofdstuk 5 is gewijd aan eenzelfde overzicht als dat in Hoofdstuk 4, maar in dit geval worden de raakvlakken van verplicht oligomere eiwitten bestudeerd. Trends betreffende de aminozuursamenstelling en -voorkeur en grootteverdeling worden beschreven. De grootste of op één na grootste patch op het gehele oppervlak van een eiwit-subunit blijkt betrokken te zijn bij het multimere raakvlak in 90% van de gevallen, hetgeen in overeenstemming is met het feit dat raakvlakken gewoonlijk meer hydrofoob zijn dan het overige eiwitoppervlak. Echter, de patches zijn niet complementair: de patches op associerende subunits vertonen geen voorkeur voor overlap. Ofschoon misschien verrassend moet dit ook verwacht worden, aangezien de vrije energie van de subunit-associatie wat betreft de patches vooral tot stand komt door het vermijden van contact van apolair oppervlak met water, en niet door het aangaan van hydrofoob contact.

Om inzicht te krijgen in het dynamisch gedrag van patches werd QUILT toegepast op moleculaire dynamica simulaties van drie verschillende eiwitstructuren. Dit is het onderwerp Hoofdstuk 6. De analyse hiervan vereist extra methoden om de QUILT patches in de tijd te volgen, en deze worden eveneens beschreven. De zgn. *patch run*s die hieruit voortvloeien laten zien dat de oppervlaktefluctuaties aanzienlijk zijn, ongeveer 25%. De meest voorkomende gemiddelde patchgrootte is ongeveer 50 $Å^2$, maar kan oplopen tot zo'n 400 $Å^2$. Een ononderbroken patch run duurt maximaal ongeveer 150 picoseconden, maar is als gevolg van de beweeglijkheid van eiwitten over het algemeen veel korter, ca. 4 picoseconden. Er is geen duidelijk verband tussen de levensduur van patch runs en hun gemiddelde grootte; wel hebben langlevende patch runs geringere fluctuaties. Hoewel het formalisme het niet uitsluit, blijken patches niet over het eiwitoppervlak te "zwerven". Dit geeft aan dat patches echt kenmerken van het eiwitoppervlak zijn. Door het clusteren van de patch runs worden de werkelijk persistente patches gevonden; ze worden *recurrent patches* genoemd. Slechts zo'n 25% hiervan heeft een sterke "liveness", dat wil zeggen, is het merendeel van de tijd vertegenwoordigd door een echte patch run. In amicyanine wijst de methode een patch aan waarvan bekend is dat hij betrokken is bij de binding van methylamine dehydrogenase. In fosfolipase A_2 wordt een grote persistente patch gevonden die bestaat uit Leu58 and Phe94. Het waarschijnlijk functionele belang hiervan was tot nog toe onbekend.

Hoofstuk 7 besluit met een bespreking van de merites van QUILT, het verband met andere methoden, alsook van resulten van andere onderzoekers. Tot slot worden aanbevelingen gedaan voor vervolgonderzoek.

Philip Lijnzaad werd geboren op 1 februari in Rotterdam. In 1983 behaalde hij zijn VWO eindexamen aan het Murmellius Gymnasium te Almaar. Van 1983 tot 1990 studeerde hij Moleculaire Wetenschappen (fysisch-chemische oriëntatie) aan de toenmalige Landbouwhogeschool in Wageningen. De afstudeeropdrachten betroffen de toepassing van een statistisch-thermodynamisch model voor water (begeleid door K. Besseling, Fysische Chemie, Wageningen), en het oplossen van een eiwitstructuur met behulp van NMR and Moleculaire Dynamica (begeleid door C.P.M. van Mierlo, Biochemie, Wageningen; en door J. de Vlieg, Fysische Chemie, Groningen). Voor laatstgenoemd werk ontving hij in 1990 de Unilever Research Prijs. Van 1990 tot 1995 was hij verbonden aan het Europees Moleculair Biologisch Laboratorium (EMBL; Heidelberg, Duitsland) in de groep van P. Argos. Daar deed hij promotieonderzoek naar hydrofobe gebieden op eiwitoppervlakken; dit werk vormt een groot deel van dit proefschrift. Van 1995 tot 2001 was hij werkzaam aan het Europees Bioinformatica Instituut (EBI; Cambridge, Verenigd Koninkrijk). Daar was hij betrokken bij de Radiation Hybrid Database (P. Rodriguez-Tomé), en bij het EnsEMBL project (E. Birney). Sinds 2001 is hij als bioinformaticus verbonden aan de afdeling Fysiologische Chemie (UMC Utrecht), onder leiding van prof. dr. F.C.P. Holstege. Een deel van het in dit proefschrift beschreven werk is hier verricht.

Philip Lijnzaad was born on February 1st, 1965, in Rotterdam, The Netherlands. In 1983, he passed his exams at Murmellius Grammar School Alkmaar. From 1983 to 1990, he read Molecular Sciences (specializing in physical chemistry) at Wageningen Agricultural University. Graduation projects were the application of a statistical-thermodynamical water model (supervised by K. Besseling, Physical Chemistry, Wageningen), and the determination of a protein structure using NMR and Molecular Dynamics (supervised by C.P.M. van Mierlo, Biochemistry, Wageningen; and by J. de Vlieg, Physical Chemistry, Groningen). For the latter, he received the 1990 Unilever Research Award. From 1990 to 1995 he worked at the European Molecular Biology Laboratory (EMBL; Heidelberg, Germany), in the P. Argos group. There, he did PhD research into hydrophobic patches on protein surfaces. This work forms an important part of this thesis. From 1995 till 2001 he was employed at the European Bioinformatics Institute (EBI; Cambridge, United Kingdom). There, he was involved with the Radiation Hybrid Database (P.

Rodriguez-Tomé), and with the EnsEMBL project (E. Birney). Since 2001, he is a bioinformatician in the department of Physiological Chemistry (University Medical Centre, Utrecht, The Netherlands), under the supervision of Prof. Dr. F.C.P. Holstege. Part of the work described in this thesis was done here.

BIBLIOGRAPHY

[1] C. P. M. van Mierlo, P. Lijnzaad, J. Vervoort, F. Müller, H.J.C. Berendsen, and J. de Vlieg. Tertiary structure of two-electron reduced megasphaera elsdenii flavodoxin and some implications, as determined by two-dimensional ^{1}H-NMR and restrained molecular dynamics. *Eur. J. Bioch.*, 194:185–98, 1990.

[2] F. Eisenhaber, P. Lijnzaad, P. Argos, C. Sander, and M. Scharf. The double cubic lattice method: Efficient approaches to numerical integration of surface area and volume, and to dot surface contouring of molecular assemblies. *Journal of Computational Chemistry*, 16:273–284, 1995.

[3] P. Lijnzaad, H.J.C. Berendsen, and P. Argos. A method for detecting hydrophobic patches on protein surfaces. *Proteins*, 26:192–203, 1996.

[4] P. Lijnzaad, H.J.C. Berendsen, and P. Argos. Hydrophobic patches on the surfaces of protein structures. *Proteins*, 25:389–397, 1996.

[5] P. Lijnzaad and P. Argos. Hydrophobic patches on protein subunit interfaces: characteristics and prediction. *Proteins*, 28:333–343, 1997.

[6] P. Rodriguez-Tomé and P. Lijnzaad. The Radiation Hybrid Database. *Nucleic Acids Research*, 25:81–84, 1997.

[7] P. Lijnzaad, J. Coppieters, T. Flores, C. Helgesen, and T. Slidel. CORBA and molecular biology. In *CORBA and Molecular Biology, a workshop of OOPSLA 1997*, 1997.

[8] P. Rodriguez-Tomé, C. Helgesen, P. Lijnzaad, and K. Jungfer. A CORBA server for the radiation hybrid database. In *Proceedings of the International Conference on Intelligent Systems and Molecular Biology*, volume 5, pages 250–253, 1997.

[9] P. Rodriguez-Tomé, C. Helgesen, P. Lijnzaad, and K. Kim Jungfer. A CORBA environment for RH maps. In *Proceedings of the Human Genome Conference*, 1997.

[10] P. Rodriguez-Tomé, C. Helgesen, and P. Lijnzaad. The Radiation Hybrid Database. In *Proceedings of the Human Genome Conference*, 1997.

[11] P. Rodriguez-Tomé, C. Helgesen, and P. Lijnzaad. The Radiation Hybrid Database. In *Proceedings of the International Genome Sequencing and Analysis Conference*, 1997.

[12] P. Rodriguez-Tomé, C. Helgesen, and P. Lijnzaad. A CORBA server for the radiation hybrid database. In *Proceedings of the German Conference on Bioinformatics*, 1997.

[13] P. Rodriguez-Tomé, C. Helgesen, and P. Lijnzaad. The Radiation Hybrid Database: Development and Services. In *First Annual Conference on Computational Genomics*, 1997.

[14] P. Lijnzaad, C. Helgesen, and P. Rodriguez-Tomé. The Radiation Hybrid Database. *Nucleic Acids Research*, 26:102–105, 1998.

[15] N. Redaschi, K. Kruszewska, P. Lijnzaad, B. Marx, P. McNeil, T. Slidel, and P. Rodriguez-Tomé. A CORBA server for the EMBL nucleotide sequence database. In *Proceedings of Objects in Bioinformatics*, 1998.

[16] E. Barillot, C. Cussat-Blanc, F. Guyon, G. Vaysseix, K. Jungfer, P. Lijnzaad, and P. Rodriguez-Tomé. A standard IDL for genome mapping. In *Proceedings of the Human Genome Conference*, 1998.

[17] P. Deloukas, G. D. Schuler, G. Gyapay, E. M. Beasley, C. Soderlund, L. P. Rodriguez-Tomé, Hui, T. C. Matise, K. B. McKusick, J. S. Beckmann, S. Bentolila, M.-T. Bihoreau, B. B. Birren, J. Browne, A. Butler, A. B. Castle, N. Chiannilkulchai, C. Clee, P. J. R. Day, A. Dehejia, T. Dibling, N. Drouot, S. Duprat, C. Fizames, S. Fox, S. Gelling, L. Green, P. Harrison, R. Hocking, E. Holloway, S. Hunt, S. Keil, P. Lijnzaad, J. C. Louis-Dit-Sully, Ma, A. Mendis, J. Miller, J. Morissette, D. Muselet, H. C. Nusbaum, A. Peck, S. Rozen, D. Simon, D. K. Slonim, R. Staples, L. D. Stein, E. A. Stewart, M. A. Suchard, T. Thangarajah, C. N. Vega-Czarny, Webber, X. Wu, J. Hudson, C. Auffray, N. Nomura, J. M. Sikela, M. H. Polymeropoulos, M. R. James, E. S. Lander, T. J. Hudson, R. M. Myers, D. R. Cox, J. Weissenbach, M. S. Boguski, and D. R. Bentley. A physical map of 30,000 human genes. *Science*, 282:44–746, 1998.

[18] N. Redaschi, K. Kruszewska, P. Lijnzaad, and P. Rodriguez-Tomé. Accessing the EMBL database through CORBA — implementation of a browsing server (EMCORBA v2). In *Proceedings of the German Conference on Bioinformatics*, 1998.

[19] E. Barillot, U. Leser, P. Lijnzaad, C. Cussat-Blanc, K. Jungfer, F. Guyon, G. Vaysseix, C. Helgesen, and P. Rodriguez-Tomé. A proposal for a standard CORBA interface for genome maps. *Bioinformatics*, 15:157–169, 1999.

[20] P. Rodriguez-Tomé and P. Lijnzaad. The Radiation Hybrid Database. *Nucleic Acids Research*, 27:115–118, 1999.

[21] P. Rodriguez-Tomé, K. Jungfer, J. Muilu, N. Redaschi, A. Robinson, M. Senger, J. Sengerova, A. Spiridou, L. Wang, and P. Lijnzaad. CORBA servers and services at EBI. In *Proceedings of the International Genome Sequencing and Analysis Conference*, 1999.

[22] Object Management Group. Biomolecular Sequence Analysis. Technical report, Object Management Group, 2000.

[23] P. Rodriguez-Tomé and P. Lijnzaad. The Radiation Hybrid Database. *Nucleic Acids Research*, 28:146–147, 2000.

[24] L. Wang, P. Rodriguez-Tomé, N. Redaschi, P. McNeil, A. Robinson, and P. Lijnzaad. Accessing and distributing EMBL data using CORBA. *Gen. Biol.*, 1:0010.1–0010.10, 2000.

[25] P. Rodriguez-Tomé and P. Lijnzaad. The Radiation Hybrid Database. *Nucleic Acids Research*, 28:165–166, 2001.

[26] L. Wang, J.J.M. Riethoven, P. Lijnzaad, N. Redaschi, and A.J. Robinson. Exploiting XML with CORBA to improve distributing EMBL data. In *Proceedings of the World Multi-Conference on Systemics, Cybernetics and Informatics*, volume X, pages 504–510, 2001.

[27] Object Management Group. Genomic Maps. Technical report, Object Management Group, 2001.

[28] T. Hubbard, D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond, L. Huminiecki, A. Kasprzyk, H. Lehvaslaiho, P. Lijnzaad, C. Melsopp, E. Mongin, R. Pettett, M. Pocock, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater, J. Smith, W. Spooner, A. Stabenau, J. Stalker, E. Stupka, A. Ureta-Vidal, I. Vastrik, and M. Clamp. The EnsEMBL genome database project. *Nucleic Acids Research*, 30:38–41, 2002. http://www.ensembl.org.

[29] M. Radonjic, J.-C. Andrau, P. Lijnzaad, P. Kemmeren, T.T. Kockelkorn, D. van Leenen, N.L. van Berkum, and F.C. Holstege. Genome-wide analyses reveal RNA polymerase II located upstream of genes poised for rapid response upon S. cerevisiae stationary phase exit. *Molecular Cell*, 18:171–183, 2005.

[30] P. Roepman, L.F. Wessels, N. Kettelarij, P. Kemmeren, A.J. Miles, P. Lijnzaad, M.G. Tilanus, R. Koole, G.J. Hordijk, P.C. van der Vliet, M.J. Reinders, P.J. Slootweg, and F.C. Holstege. An expression profile for diagnosis of lymph node metastases from primary head and neck squamous cell carcinomas. *Nat. Gen.*, 37:182–186, 2005.

[31] J.-C. Andrau, L. van de Pasch, P. Lijnzaad, T. Bijma, M. Groot Koerkamp, J. van de Peppel, M. Werner, and F.C. Holstege. Genome-wide location of the coactivator Mediator: binding without activation and transient Cdk8 interaction on DNA. *Molecular Cell*, 22:179–192, 2006.

DANK - THANKS - DANKE

Tja, dat was het dan: 16 jaar, 5 maanden, 3 weken en 5 dagen nadat ik in Heidelberg begon aan mijn promotieonderzoek EST het dan eindelijk HORA. In die lange tijd is de lijst van mensen die er direct of indirect aan hebben bijgedragen nogal gegroeid.

In de eerste plaats wil ik mijn promotoren, Frank en Jaap, bedanken voor het vertrouwen dat ze in mij stelden dat het toch allemaal goed zou komen. Frank, jou wil ik daarbij in het bijzonder bedanken voor je interesse en steun, maar vooral voor de tijd die je me gegeven hebt om het proefschrift af te maken. Ik weet niet of het zonder dat gelukt zou zijn.

"Heidelberg" is long ago, but I have very good memories of my time there in the Argos group. Pat, I would like to thank you for having me in your group, even though at the time I did not manage to obtain my PhD. Frank, unsere Diskussionen über die Boltzmann Interpretation stehen mir noch lebendig vor Augen (und ich bin noch immer der selben Meinung :-). Ich wünsche dir alles gute. André: kopje koffie? Of hebben ze dat niet in Finland? Het ga jou en Ata goed. Jaap: jou heb ik net in een andere rol bedankt, maar in Heidelberg was je misschien nog wel belangerijker voor me. Ik heb veel van je geleerd, en zonder jouw en Hellen's gezelligheid en humor zou het in Heidelberg aanmerkelijk minder leuk zijn geweest. Ben benieuwd of jullie ons misschien straks weer achterna komen verhuizen :-) Simon and Siân, it was so nice to have befriended the two of you. Pity we didn't manage to meet up more often when we lived in the UK. But there's time. Michael, es hat mir immer viel Spass gemacht mit dir zu reden über computing; du hast mich eingeführt ins Unix Geektum. Alles Gute!

Herman Berendsen was in eerste instantie betrokken als promotor. Hem ben ik erkentelijk voor de altijd verhelderende *brain storm* sessies. David van der Spoel en Pieter Meulenhoff worden bedankt voor het beschikbaar stellen van een aantal Moleculaire Dynamica trajectorieën.

Patricia, I've been lucky to have had you as a boss at the EBI. Not only were you nearly part of our extended family, you were also a technology leader. Pity CORBA didn't work out, but *c'est la vie*. Too bad Sardinia is so far away. Carsten, I have fond memories of our time together at the EBI and during our holiday in Norway. We should try to go and see some folk music together! Martin, we didn't always agree, but I always enjoyed our discussions, and wish you the best of luck outside the EBI. Rodrigo, your humour and warmth makes you very pleasurable

company, espéro que todos is bien. Wolfgang, es war super dich kennengelernt zu haben, schade dass man sich dann wieder so schnell und weit trennen muss. Sergio, we really should have jammed together more often, it was really nice making music with you. Ewan, I really enjoyed being in the EnsEMBL team, and I'm sure you'll keep up your excellent work. Jon, knowing you is special, and we were (mostly) on the same wavelength. Inviting me over for booze, a spliff and some Zappa at 01:00 AM while working on my PhD was sometimes, but not always, a good idea.

Marjoleine, bedankt voor je hulp bij de organisatie van de retraite en je inzet bij het begin van het promotietraject. In Noeline's zorgzame handen is het tot een goed einde gekomen, waarvan acte. Joop, het was erg leuk om een oud-I9 bewoner als collega te hebben. En nu dus terug in het dorp; we moesten maar eens op stap. Jeroen, de rookpauzes met jou op het Stratenumbordes waren altijd genoeglijk en leerzaam. Marijana, onze samenwerking in het laatste jaar van je promotie is tot een vriendschap geworden, al was daar de afgelopen maanden geen tijd voor. Als het goed is zijn jullie inmiddels *eindelijk* wezen eten. Jean-Christophe, it was a pleasure working with you. Your amount of biological knowledge is hard to take in for a mere bioinformatician! I'm sure you'll be doing well in Marseille. Tony, hoewel we in het werk weinig met elkaar te maken hebben stel ik je gezelschap zeer op prijs en bewonder je kennis van laboratoriumtechnieken. Paul, jij hebt je inmiddels omgeschoold (opgewerkt?) tot bioinformaticus en `R`-wizzard; super. Jammer dat we geen gebruik van je kunnen maken. Arnaud, je was een gezellige kamergenoot. Ik mis je Noord-Hollandse tongval en boze telefonades (jij waarschijnlijk niet). Veel success bij ECN! Patrick, jij bent een belangrijke reden voor het plezier in mijn werk. Zonder de vele diepgaande discussies met jou over wetenschappelijke, technische en andere zaken zou het een stuk saaier zijn. Ik heb erg veel van je geleerd (en die weddenschappen heb je met vlag en wimpel gewonnen). Harm, ook jouw gezelschap, grote kennis van zowel lab- als computerwerk en natuurlijk je cynische humor droegen enorm bij aan de goede werksfeer. Kom terug! Linda, ik hoop dat je lol hebt (en wat opsteekt) tijdens je stage. Erik, als kamergenoot en mede-bioinformaticus ben je top. Kom je straks weer naast me zitten? Ik zal niet klagen over die muziek :-) Yumas, ik hoop dat je nog wat langer in de groep blijft. Ik waardeer je humor en brede interesse, en ben ook zeer benieuwd naar je boek. Het mijne is af! Marian, je bent een ontzettend fijne collega en halve bioinformaticus. We gaan binnenkort eindelijk die cursus omgooien. Diane, we hebben niet zo veel met elkaar te maken, maar hoe mooier jullie slides, hoe makkelijker ons werk, in theorie althans. Ga zo door. Dik, ik benijd je handigheid met de robots. Heb je er al één voor je apparatenmuseum? Joris, ik ben onder de indruk van je kennis van zaken, zeker wat betreft PDF en proteomica. Maar echte mannen gebruiken natuurlijk Postgres en Perl :-) Nynke, Murmellius *Mediator bitch*, blijf nog maar even in de groep, het zou sociaal en inhoudelijk een aderlating zijn als je verdween. Eva, ik hoop dat je dit kunt lezen. Zo niet, dan moet je toch een beetje beter integreren in Nederland. Maar dat zal wel lukken, evenals

je promotie; veel succes ermee. Dat geldt ook voor jou, Loes. Sake, ik geniet ontzettend van het gezamenlijke zingen, en verheug me al op jouw promotie. Het wordt trouwens weer tijd voor een rit in de *sing-along bus*. Tineke, je mag graag mee, maar misschien wil je niet. Mariel en Nathalie, jullie moeten dit verschijnsel ook maar eens ondergaan, het is goed voor het groepsproces, al hebben jullie dat niet nodig.

Bert en Carien, jullie kennen me eigenlijk niet anders dan als promovendus. Hopelijk raken jullie nu niet uitgekeken op me; ik in ieder geval zeker niet op jullie. Beste "moleculen", het is me dan toch nog gelukt. Ik verwacht natuurlijk wel een driewerf **hoera hoer hoera, hoera hoera hoera, hoera hoera hoera**. Minimaal. Mim, Gert, Michiel, Ellis, Frank, Marjel, Martina en Axel, jullie kennen me langer promoverend dan niet. Ben benieuwd of het scheelt. Hopelijk komt er nu wat meer ruimte en lucht in mijn tijd; jullie vriendschap is me dierbaar, en we moesten maar weer eens gaan wandelen of tjalken of erger.

Jaap en Tine, ik ben jullie dankbaar voor de steun en interesse die jullie altijd getoond hebben voor dit "project", maar ook voor de ruimte die jullie ons laten. Sander en Machteld, jullie hebben een veel verstandiger volgorde aangehouden dan ik: eerst promoveren, dan kinderen. Ik had het zelf kunnen bedenken. Fijn dat jullie niet meer zo ver weg zitten. Francijn en Erik, ook jullie wil ik danken voor alle ondersteuning, gezelligheid en zorg waarmee jullie ons altijd omringen; dat maakt het een stuk makkelijker.

Beth en Liesbeth, ook van jullie weet ik dat jullie klaar stonden wanneer het nodig was; dat idee alleen al hielp, en wordt zeer gewaardeerd. Dirk en Yvonne, en Noortje en Lotte, ook jullie morele hulp-op-afstand hielp, maar het is wel erg leuk dat jullie nu dichterbij zitten. Moeder's! Wie had dat nou nog gedacht? Toch nog, ten lange leste. Dank je wel voor alles. Het is verdrietig dat Papa het niet kan meemaken. Hij zou wel trots zijn geweest, vermoed ik. Wees jij dan maar dubbel trots, in zijn plaats.

Tot slot, Dorien, en Maaike, en Fleur. Het heeft allemaal wel erg lang geduurd, en ik was niet altijd even gezellig als ik weer eens lang had zitten werken. Maar nu is dat allemaal over, als het goed is – daar mogen jullie me aan houden! Want jullie en Mama zijn voor mij het allerbelangrijkste op de hele wereld. Echt waar.

Lieve, lieve, liefste Catrien. Eindelijk is het dan echt, helemaal, voorgoed KLAAR. Waarschijnlijk ben jij nog meer opgelucht dan ik. Ik wil je uit de grond van mijn hart danken voor al je steun, al je geduld, het voor lief nemen van mijn soms onmogelijke werkuren en ditto humeur. Zonder jou was ik nergens geweest, terwijl ik soms voor jou nergens was. Dit alles overdenkend voelt het als of we (weer) gaan trouwen. Zou ik geen enkel bezwaar tegen hebben. Kus.

Zeist, Maart 2007 *Philip Lijnzaad*