Analysis of accessible surface of residues in proteins

LAURENCE LINS, ANNICK THOMAS, AND ROBERT BRASSEUR

Centre de Biophysique Moléculaire Numérique (CBMN), Faculté des Sciences Agronomiques de Gembloux (FSAGx), 5030 Gembloux, Belgium

(RECEIVED January 31, 2003; FINAL REVISION March 26, 2003; ACCEPTED March 28, 2003)

Abstract

We analyzed the total, hydrophobic, and hydrophilic accessible surfaces (ASAs) of residues from a nonredundant bank of 587 3D structure proteins. In an extended fold, residues are classified into three families with respect to their hydrophobicity balance. As expected, residues lose part of their solvent-accessible surface with folding but the three groups remain. The decrease of accessibility is more pronounced for hydrophobic than hydrophilic residues. Amazingly, Lysine is the residue with the largest hydrophobic accessible surface in folded structures. Our analysis points out a clear difference between the mean (other studies) and median (this study) ASA values of hydrophobic residues, which should be taken into consideration for future investigations on a protein-accessible surface, in order to improve predictions requiring ASA values. The different secondary structures correspond to different accessibility of residues. Random coils, turns, and β-structures (outside β-sheets) are the most accessible folds, with an average of 30% accessibility. The helical residues are about 20% accessible, and the difference between the hydrophobic and the hydrophilic residues illustrates the amphipathy of many helices. Residues from β-sheets are the most inaccessible to solvent (10% accessible). Hence, β-sheets are the most appropriate structures to shield the hydrophobic parts of residues from water. We also show that there is an equal balance between the hydrophobic and the hydrophilic accessible surfaces of the 3D protein surfaces irrespective of the protein size. This results in a patchwork surface of hydrophobic and hydrophilic areas, which could be important for protein interactions and/or activity.

Keywords: Solvent accessibility; Pex files; hydrophobicity; secondary structure; amphipathy

Understanding the folding of proteins remains one of the major scientific challenges. One way to explore this complex problem is to get information from the protein structures themselves. We recently developed an analytical tool, named Pex files, in which numerical data on various structural parameters of proteins are described, such as secondary structures, side chain interactions, H-bonds, and more (Thomas et al. 2001, 2002a,b). Here we introduce a new Pex file that, in addition to the major structural parameters of proteins, lists a series of parameters describing the solvent accessibility.

The folding process of soluble proteins decreases the surface in contact with the solvent. This is related to the secondary structures of proteins. Accurate knowledge of residue accessibility would thus aid the prediction of secondary structures. Different methods of prediction are based on the use of protein structure databases and on multiple sequence alignments. They have various efficiencies, notably depending on the number of relative accessibility states (i.e., exposed, buried, and in-between; Rost and Sander 1994; Rost 1996; Li and Pan 2001; Naderi-Manesh et al. 2001; Yuan et al. 2002).

Further, because active sites of proteins are often located at the surface of the protein, greater insight into residue accessibility would be important in understanding and predicting structure/function relationships.

In the present study, we analyzed 587 proteins from the Protein Data Bank (PDB) using the Pex files. We extracted the total, hydrophobic, and hydrophilic accessible surfaces of residues. The method used to calculate the accessible surface is that of Shrake and Rupley (1973). The 587-pro-

Reprint requests to: Robert Brasseur, CBMN, Passage des Déportés, 2, FSAGx, B-5030 Gembloux, Belgium; e-mail: brasseur.r@fsagx.ac.be; fax: 32-8162-2522.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.0304803.

Petitioner Merck, Ex. 1035, p. 1406

tein bank is a nonredundant bank of structures (Liu and Chou 1999).

Results

Calculation of residue accessible surfaces in extended conformation

To check that the calculation of accessible surfaces in ASA-Pex files is correct, we used a window of three residues along the structure; the surface of the central residue is calculated. This mimics the surface of residues measured in tripeptide by others, such as Gly-X-Gly or Ala-X-Ala (Creighton 1993; Samanta et al. 2002) and corresponds to the residue surface in the unfolded state. A very good correlation is observed between Pex data and previously published values (Table 1). When the residue surface is split into hydrophilic and hydrophobic surfaces (Table 1; Fig. 1), the residues with the highest hydrophobic versus hydrophilic ASA ratio are Phe and Met. Their hydrophobic surface is more than four times higher than the hydrophilic one. Residues with the smallest ratio (3–4 times more hydrophilic than hydrophobic) are Asp and Asn. Plotting the hydrophobic/hydrophilic ratio as a function of the total ASA value clusters the residues into three groups (Fig.1): One group is the hydrophobic amino acids (G<A<V,C,P<I,L<F,M) for which the hydrophobic/hydrophilic ratio is increasing with the surface; the second contains the hydrophilic residues (D,N<E,Q<R) whose ratio is rather independent of the residue surface; and a third group containing S,T,H,K,Y and W, whose ratio varies almost exponentially with the residue surface. This underlines that the aromatic residues with polar atoms (W,Y,H) do not behave like Phenylalanine and are not pure hydrophobic residues. It is worth noting that Lysine, although generally considered as a hydrophilic residue, has a ratio near 1 (Fig.1) and belongs to the third group, unlike Arg. This is due to its long hydrophobic chain holding the polar head.

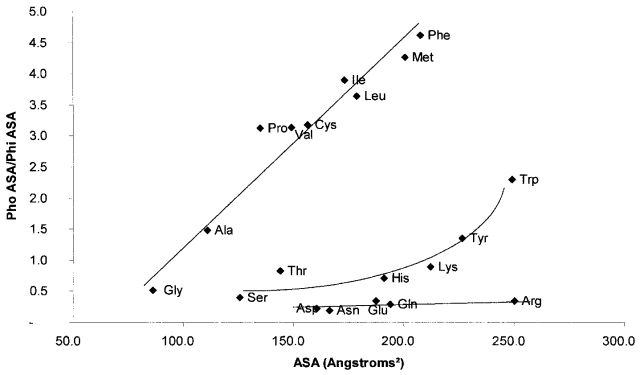

Figure 1. Hydrophobic/hydrophilic accessible surface ratio (corresponding to column 6 of Table 1) as a function of the total accessible surface for each residue (Table 1, column 3) in the unfolded state.

Table 1. *Total, hydrophobic (pho), and hydrophilic (phi) accessible surfaces of whole residues (backbone and lateral chain) calculated with a window 3*

Residue name	Surface Creighton	Total accessible surface	Hydrophobic (pho) surface	Hydrophilic (phi) surface	Ratio pho surface/phi surface
arg	241	250	65	187	0.3
trp	259	249	174	76	2.3
tyr	229	227	130	96	1.4
lys	211	212	101	114	0.9
phe	218	208	170	37	4.6
met	204	201	163	38	4.3
gln	189	194	45	152	0.3
his	194	191	79	111	0.7
glu	183	187	49	141	0.3
leu	180	179	140	39	3.6
ile	182	173	137	35	3.9
asn	158	166	27	140	0.2
asp	151	160	30	131	0.2
cys	140	157	119	38	3.2
val	160	149	112	36	3.1
thr	146	144	65	79	0.8
pro	143	135	103	33	3.1
ser	122	125	36	89	0.4
ala	113	111	66	45	1.5
gly	85	86	29	56	0.5

Surfaces are expressed in angströms^2. Reference surfaces are those determined experimentally on Gly-X-Gly peptides by Creighton (1993). The ratio of pho surface vs. phi surface corresponds to the ratio of column 4 vs. column 5.

Petitioner Merck, Ex. 1035, p. 1407

Calculation of residue accessible surface in folded proteins

Accessible surfaces of residues in folded proteins were determined as described in Materials and Methods. As expected, all residues decrease their accessibility to the solvent, to about 45% for Lys and Glu and down to a few percent (about 5%) for the most hydrophobic residues (Ile, Leu, Val, and Cys), with respect to the residue accessibility of the unfolded state (Table 2). On average, the accessible surface of residues of folded proteins is reduced to 20%.

There is a clear difference in the behavior of the hydrophobic and hydrophilic residues, the latter being largely more accessible, as shown in Figure 2. Folding similarly reduces the hydrophobic and hydrophilic surfaces for each residue, except for Proline (and to a lesser extent Glycine), which buries much more hydrophilic than hydrophobic surface (Table 2, Fig. 2). Figure 3 summarizes these observations, showing the hydrophilic versus the hydrophobic accessible surface of the residues of folded proteins. The same three groups of residues observed in Figure 1 can be distinguished. They correspond to the hydrophobic residues (C,I,L,V,A,M,F,G with low hydrophobic and hydrophilic ASAs) and to the hydrophilic residues (N,D,Q,E,R with high hydrophilic and low hydrophobic ASAs); and the third group (S,T,Y,W,H) has intermediate hydrophobic and hydrophilic ASA values. Note the peculiar behavior of Pro and Lys in that both have high hydrophobic ASA but differ in their hydrophilic ASA value.

Values of the accessible surface of hydrophilic residues are in good agreement with those previously reported (Samanta et al. 2002). However, the ASA for the hydrophobic amino acids and especially for Ileu, Leu, Val, Cys, and Met are 2 to 3 times lower than the previously reported values (Table 2). This is due to the fact that we used median instead of mean ASA values. We believe that median ASA values are more appropriate than means when we look at the frequency distribution of residue ASAs in 3D structures (Fig. 4). Where the ASA distribution of hydrophilic residues such as Lys or Gln appears partly gaussian, indicating that mean and median values should not be too different (Fig.4A,B, respectively), frequency distributions of the hydrophobic residues ASA are clearly not symmetrical. Many of those amino acids are completely inaccessible to the solvent, as shown for Ile (Fig. 4C) or Phe (Fig. 4D). Mean values of these distributions correspond to the previously reported ASA values of amino acids as shown in Table 2, suggesting that our data are not different from the literature. However, mean values are obviously inadequate to describe the distribution of ASAs, the median being more appropriate. This suggests that ASAs of hydrophobic residues were overestimated in a number of studies.

Accessible surfaces in the different secondary structures

Five classes of secondary structures were considered: α-helix (Ha), parallel (Bp) and antiparallel (Ba) β-strands,

Table 2. *Median values of total, hydrophobic (pho), and hydrophilic (phi) accessible surface (ASA) of whole residues from folded proteins*

Residue name	Total median ASA	Hydrophobic median ASA	Hydrophilic median ASA	% total ASA vs win3	% pho ASA vs. win3	% phi ASA vs. win3	Total mean ASA from Samanta	Total mean ASA (this study)
ala	14	7	3	12	10	7	28	27
arg	87	20	65	35	31	35	86	91
asn	59	8	49	36	29	35	58	62
asp	62	11	49	39	36	38	58	63
cys	5	2	0	3	1	0	17	15
gln	74	15	56	38	34	37	69	75
glu	83	20	60	44	41	42	73	81
gly	19	8	8	22	27	14	27	26
his	46	20	26	24	25	23	54	56
ile	6	4	0	4	3	0	25	23
leu	9	5	0	5	4	1	29	26
lys	102	47	54	48	47	47	96	101
met	13	9	1	7	6	2	36	36
phe	13	8	0	6	5	0	31	31
pro	49	40	5	36	39	15	54	51
ser	35	10	20	28	28	23	39	41
thr	37	17	15	26	26	19	44	44
trp	25	14	7	10	8	9	44	42
tyr	31	12	15	14	10	15	46	43
val	8	5	0	5	4	1	24	23

Percentages are obtained from the win 3 values (Table 1).
Surfaces (mean and median values) are compared to the mean values published in Samanta et al. 2002.
Surfaces are expressed in angströms^2.

Petitioner Merck, Ex. 1035, p. 1408

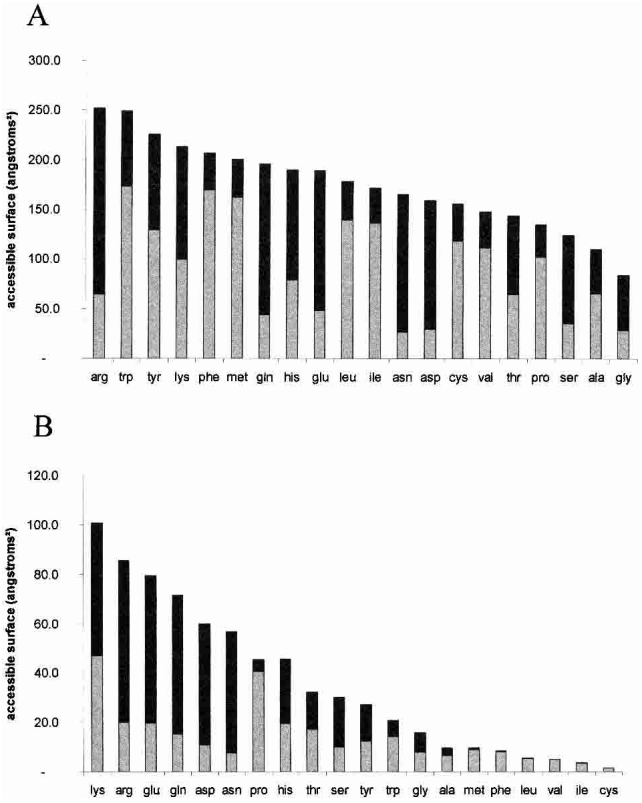

Figure 2. Values of median pho (black bars) and phi (gray bars) ASA for each residue in the extended state (*A*) and in folded proteins (*B*). Residues are sorted by decreasing total ASA values.

β-structures (B), and random coil/turn (C-T) conformations. These structural elements are defined in Materials and Methods. They correspond to 32,806 residues analyzed for Ha; 2851 residues for Bp; 27,603 residues for Ba; 36,230 residues for B; and 45,220 residues for C-T.

For each structural class, the total, hydrophobic, and hydrophilic median ASAs were calculated (Tables 3–7) and compared to the residue ASA in the extended conformation (ASA calculated with a window of 3). The most accessible residues belong to the random coil/turn (C-T) class, whereas the Ba and Bp structures result in the most solvent-inaccessible residues.

Random coil/turn structures

This class has the most accessible residues, with an average of 30% of accessible surface (Table 3). There is a segregation between the hydrophobic (I,L,V,M,W,F,Y) and the hydrophilic (K,R,N,D,Q,E,T,S) residues. The former are 10%–20% percent accessible and the latter are about 50% accessible. Gly, His, and Ala have intermediate values, and Pro is highly accessible. Hydrophobic and hydrophilic accessible surfaces are similarly decreased, except for Gly and Pro.

It should be noted that the hydrophilic accessible surface of the most hydrophobic residues (i.e., Ile, Val, Leu, Met, Phe) should correspond to the accessibility of their backbone (Table 3).

β-Strands (B-structures)

The residues are almost as accessible in B-structures as in random coil/turn structures, and the segregation between the hydrophobic and the hydrophilic residues is also observed (Table 4). A noticeable difference lies in how the hydrophobic and hydrophilic accessible surfaces of residues decrease (cf. columns 9, 10 of Tables 3 and 4). Although the hydrophobic and hydrophilic ASAs of hydrophilic residues are similarly reduced, hydrophobic residues show a twofold more pronounced decrease in the hydrophobic surface compared to the hydrophilic surface. This suggests that B-structure is more prone to shield the hydrophobic moiety of hydrophobic amino acids compared to the random coil/turn structure.

For Proline, which is 30% accessible, the opposite is observed; that is, the hydrophilic surface is less accessible than the hydrophobic one (as in the random coil/turn structure).

Further, as for random coil/turn, the hydrophilic accessible surface of the most hydrophobic residues in B-structures should correspond to the accessibility of the backbone.

Parallel and antiparallel sheets

These folds correspond to the less accessible residues: On average, only 10% of the residue surface is accessible (Tables 5,6). This is particularly true for the hydrophobic residues such as Leu, Ile, Val, Met, Cys, and Phe (1%–5% accessibility). In these folds, again, the most accessible residues are Lys and Glu (23%–33% accessibility) and hydro-

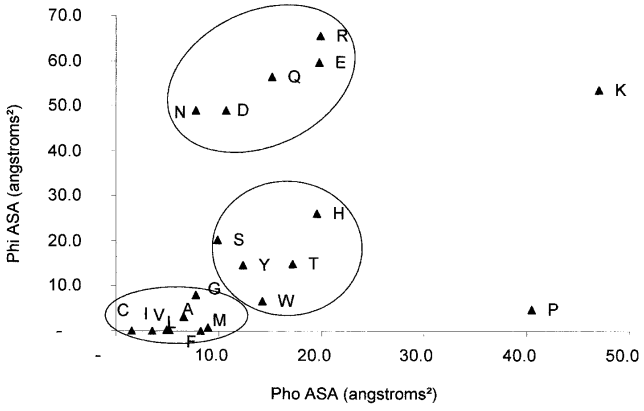

Figure 3. Median hydrophilic ASA as a function of median hydrophobic ASA of residues in folded proteins. Residues with similar behavior are grouped. Pro (P) and Lys (K) are left out of this classification.

Petitioner Merck, Ex. 1035, p. 1409

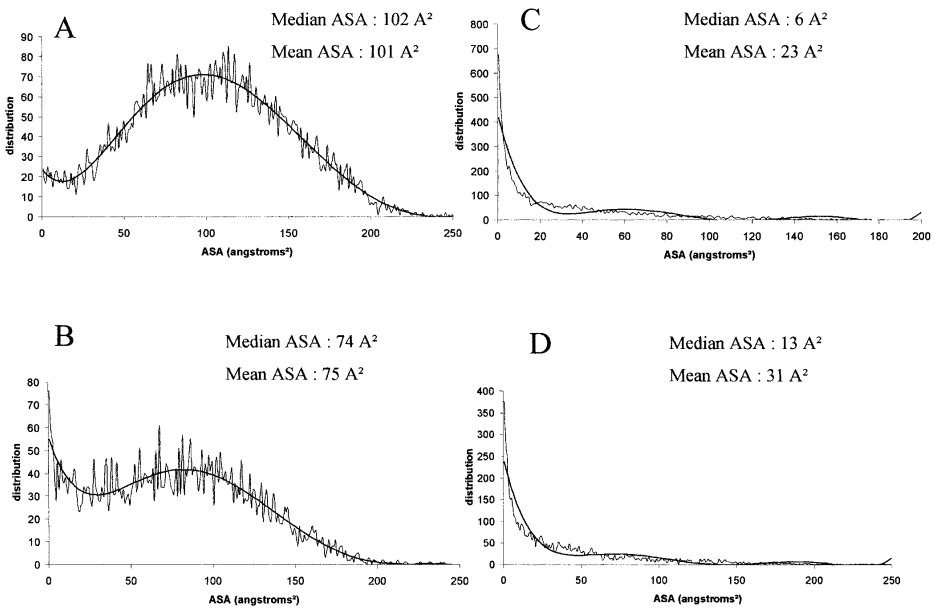

Figure 4. Distribution of the total ASA for (*A*) Lysine, (*B*) Glutamine, (*C*) Isoleucine, and (*D*) Phenylalanine. Values of the median and mean total ASA are indicated for each of these residues.

phobic and hydrophilic residues are segregated, Ser and Thr having intermediate values. For almost all residues, the Ba/Bp structures appear to shield their hydrophobic domains from the solvent better than any other structure, as reported by Chothia (1976).

This suggests that a sequence will have a smaller hydrophobic accessible surface as a β-sheet than in any other conformation. This could be related to the formation of fibrils observed with highly hydrophobic peptides. Indeed, fibrils are made of antiparallel β-strands, as reported regard-

Table 3. *Median values of total, pho, and phi ASA of whole residues in random coil/turn structures*

Residue name	Total ASA C/T	pho ASA C/T	phi ASA C/T	Total ASA win3	pho ASA win3	phi ASA win3	% total ASA	% ASA pho	% ASA phi
ala	43	27	13	111	66	45	39	41	28
arg	106	28	79	250	65	187	42	44	42
asn	81	13	67	166	27	140	48	48	48
asp	84	17	64	160	30	131	52	56	49
cys	14	7	4	157	119	38	9	6	10
gln	96	23	70	194	45	152	50	52	46
glu	105	30	70	187	49	141	56	62	50
gly	29	13	15	86	29	56	34	46	27
his	64	27	36	191	79	111	33	34	33
ile	30	20	4	173	137	35	17	15	10
leu	27	18	4	179	140	39	15	13	11
lys	121	59	64	212	101	114	57	58	56
met	51	37	7	201	163	38	25	23	18
phe	31	22	4	208	170	37	15	13	10
pro	60	48	7	135	103	33	45	47	22
ser	64	20	40	125	36	89	51	55	45
thr	58	28	29	144	65	79	40	42	36
trp	41	28	13	249	174	76	16	16	17
tyr	43	22	20	227	130	96	19	17	21
val	31	22	4	149	112	36	21	19	11

ASAs are expressed in angströms^2 and are compared to those obtained with a window 3.

Table 4. *Median values of total, pho, and phi ASA of whole residues in B structure*

Residue name	Total ASA B	pho ASA B	phi ASA B	Total ASA win3	pho ASA win3	phi ASA win3	% total ASA	% ASA pho	% ASA phi
ala	24	12	9	111	66	45	22	18	19
arg	94	20	71	250	65	187	38	32	38
asn	57	8	46	166	27	140	34	30	33
asp	62	12	47	160	30	131	38	41	36
cys	19	4	6	157	119	38	12	4	15
gln	82	15	64	194	45	152	43	34	42
glu	89	19	67	187	49	141	47	38	47
gly	16	5	7	86	29	56	19	17	13
his	55	22	31	191	79	111	29	28	28
ile	23	10	6	173	137	35	14	8	16
leu	27	12	7	179	140	39	15	9	18
lys	104	47	55	212	101	114	49	46	49
met	33	20	7	201	163	38	16	12	19
phe	33	18	7	208	170	37	16	10	18
pro	40	34	4	135	103	33	30	33	11
ser	38	12	20	125	36	89	31	33	22
thr	50	22	20	144	65	79	35	34	25
trp	38	19	18	249	174	76	15	11	24
tyr	44	19	24	227	130	96	19	14	25
val	24	11	6	149	112	36	16	10	18

ASAs are expressed in angströms^2 and are compared to those obtained with a window 3.

Petitioner Merck, Ex. 1035, p. 1410

Table 5. *Median values of total, pho, and phi ASA of whole residues in Ba structure*

Residue name	Total ASA Ba	pho ASA Ba	phi ASA Ba	Total ASA win3	pho ASA win3	phi ASA win3	% total ASA	% ASA pho	% ASA phi
ala	1	0	0	111	66	45	1	0	0
arg	59	8	47	250	65	187	24	12	25
asn	21	1	18	166	27	140	13	4	13
asp	23	2	20	160	30	131	14	7	15
cys	1	0	0	157	119	38	0	0	0
gln	38	3	33	194	45	152	20	6	22
glu	43	5	35	187	49	141	23	9	25
gly	1	0	0	86	29	56	1	0	0
his	21	10	11	191	79	111	11	13	10
ile	1	1	0	173	137	35	1	1	0
leu	2	1	0	179	140	39	1	1	0
lys	69	28	40	212	101	114	33	28	35
met	2	1	0	201	163	38	1	1	0
phe	4	2	0	208	170	37	2	1	0
ser	8	1	5	125	36	89	7	2	6
thr	14	5	6	144	65	79	9	7	7
trp	10	6	1	249	174	76	4	3	1
tyr	17	4	9	227	130	96	7	3	9
val	1	1	0	149	112	36	1	1	0

ASAs are expressed in angströms^2 and are compared to those obtained with a window 3.

ing the amyloid aggregates of Alzheimer's disease (Li et al. 1999; Schladitz et al. 1999).

It is interesting to note that the backbone of hydrophobic residues (corresponding to the phi ASA of those residues) is no more accessible in Ba/Bp structures, in contrast to what happens for random coil/turn and B-conformations.

α-*Helices*

Residues of helical folds have an intermediate solvent accessibility of ~20% (Table 7). The difference between the accessibility of hydrophobic and hydrophilic residues is highly marked. For α-helices, K, E, R, N, D, Q, S, and T are 45% (Lys) to 19% (Ser) accessible with an average of 35% accessibility, whereas hydrophobic residues are only 1%–5% accessible (except for Trp and Tyr, which are 10% accessible). This is linked to the observation that most α-helices of protein 3D structures are amphipathic (Chou et al. 1997). Amphipathic helices have most of the hydrophobic residues oriented toward the protein core, whereas the hydrophilic residues are water-accessible.

It should be noted that the helical structure similarly reduces the hydrophobic and the hydrophilic surfaces of all residues, in contrast to what happens in β-folds.

In the helical structure, as for β-sheets, the backbone is not accessible (hydrophilic surfaces of hydrophobic residues are almost null).

The same calculations of surfaces were made by selecting the secondary structures attributed to the CO-side of the residue. The number of residues analyzed for each structural class remains similar, as were the surfaces (data not shown).

Table 6. *Median values of total, pho, and phi ASA of whole residues in Bp structure*

Residue name	Total ASA Bp	pho ASA Bp	phi ASA Bp	Total ASA win3	pho ASA win3	phi ASA win3	% total ASA	% ASA pho	% ASA phi
ala	5	1	1	111	66	45	5	2	2
arg	54	8	46	250	65	187	22	12	24
asn	16	0	15	166	27	140	10	0	10
asp	28	1	25	160	30	131	17	5	19
cys	1	0	0	157	119	38	1	0	0
gln	44	5	38	194	45	152	23	11	25
glu	54	9	42	187	49	141	29	19	30
gly	4	0	1	86	29	56	5	0	1
his	25	11	13	191	79	111	13	14	12
ile	5	1	0	173	137	35	3	1	0
leu	7	2	0	179	140	39	4	2	0
lys	69	30	39	212	101	114	33	30	35
met	8	4	1	201	163	38	4	2	2
phe	11	7	0	208	170	37	5	4	0
ser	11	2	5	125	36	89	8	6	6
thr	13	4	6	144	65	79	9	6	8
trp	17	13	5	249	174	76	7	8	6
tyr	25	10	12	227	130	96	11	7	13
val	6	3	0	149	112	36	4	3	0

ASAs are expressed in angströms^2 and are compared to those obtained with a window 3.

Table 7. *Median values of total, pho, and phi ASA of whole residues in Ha structure*

Residue name	Total ASA Ha	pho ASA Ha	phi ASA Ha	Total ASA win3	pho ASA win3	phi ASA win3	% total ASA	% ASA pho	% ASA phi
ala	5	2	1	111	66	45	5	4	3
arg	86	20	64	250	65	187	34	31	34
asn	53	6	45	166	27	140	32	24	32
asp	59	9	48	160	30	131	37	32	36
cys	2	1	0	157	119	38	1	1	0
gln	69	14	53	194	45	152	36	32	35
glu	76	18	57	187	49	141	41	37	40
gly	3	1	1	86	29	56	3	3	1
his	45	19	24	191	79	111	24	25	22
ile	4	3	0	173	137	35	2	2	0
leu	4	3	0	179	140	39	2	2	0
lys	95	44	51	212	101	114	45	43	45
met	6	4	0	201	163	38	3	3	0
phe	7	5	0	208	170	37	3	3	0
ser	24	6	16	125	36	89	19	17	18
thr	26	11	10	144	65	79	21	16	13
trp	16	11	3	249	174	76	11	6	4
tyr	24	10	11	227	130	96	10	7	11
val	3	3	0	149	112	36	1	2	0

ASAs are expressed in angströms^2 and are compared to those obtained with a window 3.

Petitioner Merck, Ex. 1035, p. 1411

Analysis of ASAs in data sets containing only β- or α-proteins

In light of our observation that the β-fold better shields hydrophobic parts of amino acids from water whereas helical structure better segregates hydrophobic and hydrophilic residues (the latter remaining accessible to the solvent), we wondered whether this would also hold true for two data sets one containing proteins with α-structures (no β-residues) and the other with β-proteins (no helical folds). These sets were extracted from the 587-protein bank by selecting 26 β-proteins and 55 α-proteins, as described in Materials and Methods.

Table 8 shows that the accessible surfaces of residues in helical conformation of α-proteins (corresponding to 6,530 residues) are the same as those determined for the 587 proteins, confirming the amphipathic character of the helical fold.

In Table 9, Ba and Bp structures were grouped, because only 144 residues were from parallel β-sheets. The two structures altogether correspond to 1,273 residues. Hydrophobic residues are poorly accessible, but the ASAs of hydrophilic residues (R,N,D,E,H,K,S,T) are higher (20%–50% increase) than those in the Ba and Bp residues analyzed in the 587-protein bank. This is most likely because the β-proteins must finally be soluble: Indeed, they contain more hydrophilic residues (42% as compared to 27% for the 587-protein bank). Nonetheless, residues in β-structures remain the less accessible (about 12% on average compared to 20% for the helical structure) and again, β-sheet is the best fold to shield the hydrophobic part of amino acids from water.

Table 8. *Median values of total, pho, and phi ASA of whole residues in Ha structure extracted from the alpha proteins bank*

Residue name	Total ASA Ha	pho ASA Ha	phi ASA Ha	Total ASA win3	pho ASA win3	phi ASA win3	% total ASA	% ASA pho	% ASA phi
ala	8	5	2	111	66	45	8	8	5
arg	90	23	68	250	65	187	36	36	36
asn	57	8	49	166	27	140	34	28	35
asp	56	9	46	160	30	131	35	29	35
cys	2	1	0	157	119	38	2	1	1
gln	71	15	51	194	45	152	37	34	33
glu	71	15	51	187	49	141	38	31	36
gly	8	4	2	86	29	56	9	15	4
his	52	22	31	191	79	111	27	28	28
ile	6	5	0	173	137	35	3	4	0
leu	5	4	0	179	140	39	3	3	1
lys	92	43	50	212	101	114	43	43	44
met	6	5	0	201	163	38	3	3	1
phe	9	7	0	208	170	37	4	4	0
ser	27	8	15	125	36	89	21	23	17
thr	28	11	14	144	65	79	19	17	18
trp	22	15	4	249	174	76	9	8	5
tyr	22	8	9	227	130	96	10	6	10
val	5	4	0	149	112	36	3	4	0

ASAs are expressed in angströms^2 and are compared to those obtained with a window 3.

Analysis of the total surface of the 587 proteins

We next examined the relationships between the total accessible surface of a folded protein and its sequence length. Figure 5 shows that the water-accessible surface of proteins is a simple function of the number of residues. This is in agreement with the study of 12 proteins by Chothia (1976).

More surprising is the ratio of hydrophilic to hydrophobic accessible surfaces, which lies near 1 for all of the proteins we analyzed. Figure 6 shows that hydrophobic and hydrophilic accessible surfaces are proportional to the total protein surface (regression coefficient near 0.99) and that the proportionality is similar for hydrophobic and hydrophilic surfaces. This supports that the surface of a soluble protein is not only hydrophilic, as often meant. These data suggest that folding tends to equilibrate the hydrophobic and hydrophilic solvent-accessible areas. This is illustrated by calculating the molecular hydrophobicity potentials (MHPs; Brasseur 1991) that draw the envelope of the hydrophobic and hydrophilic environments of a molecule (Fig. 7) and demonstrates that soluble molecules have significant hydrophobic patches.

Discussion

We analyzed the water-accessible surfaces of all residues from 587 nonredundant structures of proteins determined by NMR and X-ray diffraction with 1.2 to 2.9 Å resolution. As expected, we show that residues lose part of their accessible surface due to folding and that the decrease is more pronounced for hydrophobic residues, as previously described (Chothia 1976). However, we also report that the accessible surfaces of hydrophobic amino acids in proteins are for the most part smaller than previously reported (Samanta et al. 2002). Many hydrophobic residues are completely buried, and the distributions of residue ASAs are far from being gaussian. Hence, the average accessibility of a residue in a protein is better described by a median than by a mean value. Because our analysis depicts a clear difference between the mean and median ASA values of hydrophobic residues, this should be taken into consideration for future investigations of protein accessible surfaces and would probably help to improve predictions using ASA values.

Three groups of amino acids can be distinguished based on the relationships between their hydrophobic and hydrophilic accessible surfaces, either in the extended state or in the folded proteins. One group is made of the hydrophobic residues (I,L,V,F,M,A,G), another contains the hydrophilic residues (D,N,E,Q,R), and the third shows intermediate behavior (H,Y,W,S,T). Proline and Lysine are apart. Among the hydrophilic residues, Lysine is peculiar: It is often looked at as a hydrophilic residue but in the unfolded state,

Petitioner Merck, Ex. 1035, p. 1412

Table 9. *Median values of total, pho, and phi ASA of whole residues in Ba/Bp structures extracted from the beta proteins bank*

Residue name	Total ASA Ba/Bp	pho ASA Ba/Bp	phi ASA Ba/Bp	Total ASA win3	pho ASA win3	phi ASA win3	% total ASA	% ASA pho	% ASA phi
ala	6	3	1	111	66	45	6	4	1
arg	75	9	55	250	65	187	30	15	29
asn	32	1	28	166	27	140	19	5	20
asp	42	4	30	160	30	131	26	13	23
cys	1	0	0	157	119	38	1	0	1
gln	46	5	38	194	45	152	24	10	25
glu	45	8	38	187	49	141	24	16	27
gly	3	0	1	86	29	56	4	1	2
his	45	19	24	191	79	111	24	25	22
ile	7	5	0	173	137	35	4	4	0
leu	6	4	0	179	140	39	3	3	0
lys	61	29	39	212	101	114	29	29	34
met	4	2	0	201	163	38	2	1	0
phe	6	5	0	208	170	37	3	3	1
ser	18	3	11	125	36	89	15	9	12
thr	23	8	10	144	65	79	16	12	13
trp	16	12	7	249	174	76	6	7	9
tyr	33	9	18	227	130	96	15	7	19
val	6	4	0	149	112	36	4	3	1

ASAs are expressed in angströms^2 and are compared to those obtained with a window 3.

it has almost equal hydrophobic and hydrophilic accessible surfaces. Moreover, in the folded state, it has the highest hydrophobic ASA of all residues. Proline also has special features, as it has the highest hydrophobic accessible surface among the hydrophobic residues. This is related to its preferential location in accessible turns of proteins.

The classification of amino acids into three amino acid "families" following their hydrophobic/hydrophilic accessible surface should be important in terms of the prediction of conservative mutations.

The different types of secondary structure correspond to different accessible surfaces. Random coils, turns, and β-strands that are either not H-bound or are H-bound to a structure that is not a strand are the most accessible folds, with an average of 30% of residue accessibility. In these structures, the backbone of the most hydrophobic residues (I,V,L,M,F) is quite accessible, with 10%–15% accessibility.

The β-sheets (parallel and antiparallel strands) are the most solvent-inaccessible structures (with about 10% of

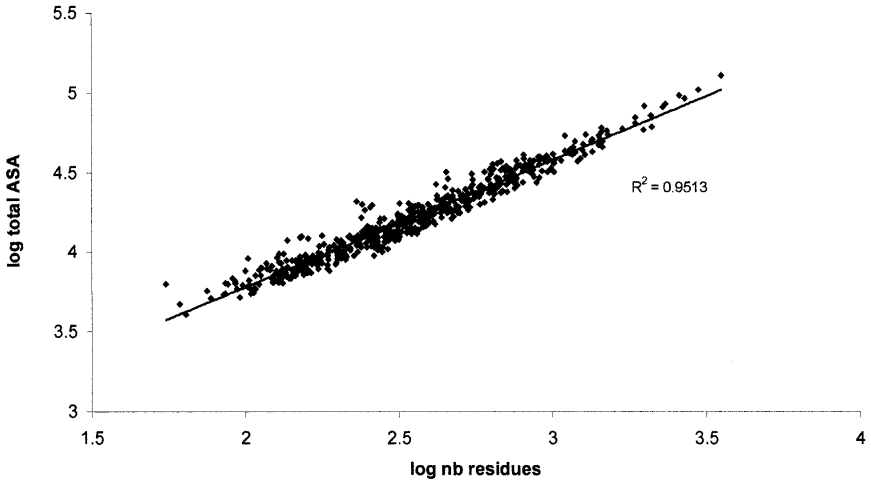

Figure 5. Log of the total ASA as a function of the log of the number of residues for the 587 proteins analyzed. The coefficient of the linear regression is 0.97.

Petitioner Merck, Ex. 1035, p. 1413

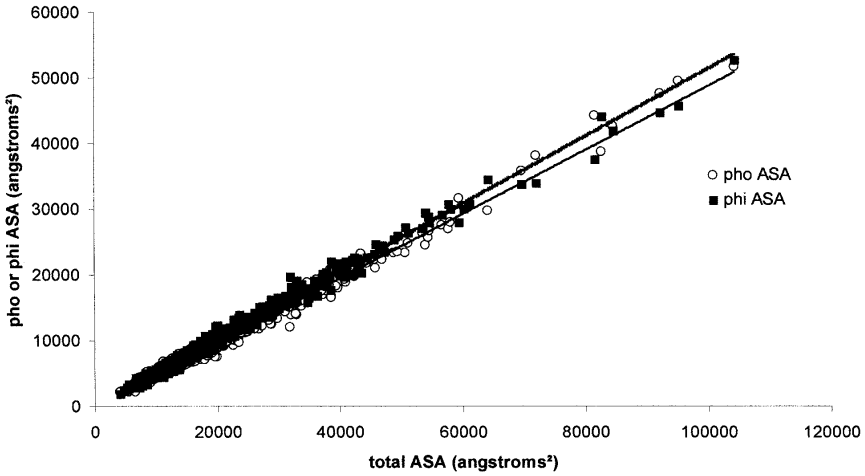

Figure 6. Hydrophobic (○) and hydrophilic (■) ASA as a function of the total ASA for the 587 proteins analyzed. The coefficient of the linear regression is 0.99 for both curves.

residue accessibility), whereas the helical conformation has an intermediate value, with about 20% of the residue surface accessible.

Both helical and β-sheet conformations shield the backbone of the most hydrophobic residues from water, in contrast to what happens for the "unordered" structures.

In all folds, there is a noticeable difference between the hydrophobic and hydrophilic residues, the latter being always more solvent-accessible. The greatest difference is observed in α-helices related to their amphipathic character. When the protein folds, the hydrophobic side of the helix is buried in the protein core, and the hydrophilic side remains solvent-accessible. The β-sheets are the most appropriate structures to shield the hydrophobicity of residues. This is likely important in the formation of fibrils in pathological and nonpathological phenomena.

Note that Lysine and Glutamic acid are the most accessible residues, whereas Leucine, Isoleucine, and Valine are the most inaccessible, irrespective of secondary structures.

As described earlier by Chothia (1976) regarding 12 proteins, there is a simple relationship between the total accessible surface of folded proteins and their size. We also show that there is a balance between the hydrophobic and the hydrophilic surfaces of the 3D protein surface. This balance is maintained irrespective of the protein size, resulting in a patchwork surface of hydrophobic and hydrophilic areas. Size and accessibility of the patches should be important for protein-protein interaction sites and/or for activity, as suggested by others (Eisenhaber and Argos 1996; Jones and Thornton 1997).

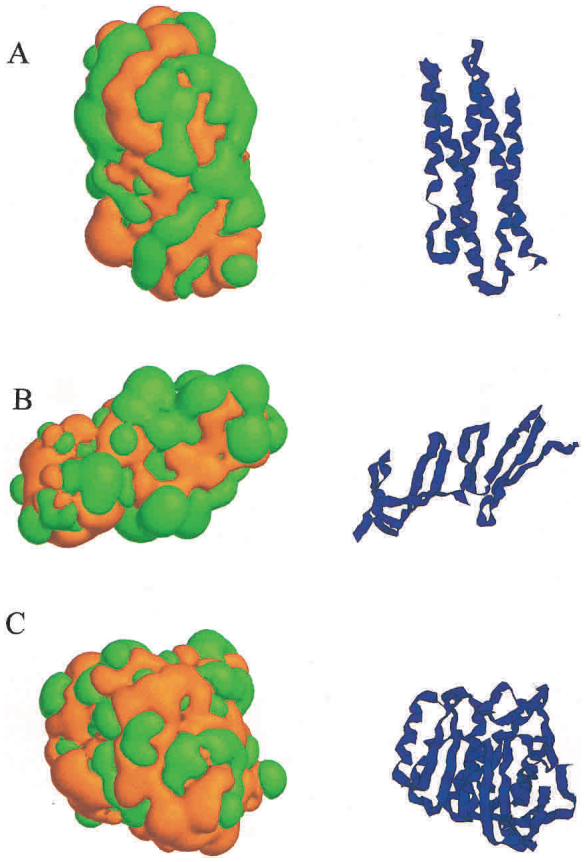

Figure 7. (*Left* panels) Molecular hydrophobicity potentials (MHPs) around three proteins analyzed (one α-protein, apolipophorin III, PDB code 1AEP; one β-protein, human fibronectin, FBR; one α/β-protein, dienelactone hydrolase, 1DIN). MHPs, based on atomic transfer energies, allow the visualization of pho (orange) and phi (green) domains around a protein and are calculated as described in (Brasseur 1991). (*Right* panels) Ribbon representations of the protein in the same orientation as in the left panels. (*A*) 1AEP, (*B*) 1FBR, (*C*) 1DIN.

Materials and methods

Proteins

The PDB files of the 587 proteins were transformed into ASA Pex files, giving a total of 156,215 residues that were analyzed. These

Petitioner Merck, Ex. 1035, p. 1414

structures were issued from the nonredundant bank described by Liu and Chou (1999) containing 593 entries. Five proteins were discarded, their format being incompatible with their transformation into Pex files. The PDB codes are listed below; the experimental resolution is between 1.2 and 2.9 Å. ASA Pex files are accessible on the CBMN Web site (http://www.fsagx.ac.be/bp/) or can be obtained from author R. Brasseur.

153l	1bec	1chd	1daa	1ecl	1ftp	1gtr
1aa8	1ber	1chk	1dar	1ecp	1fua	1gym
1aaf	1bgw	1chm	1dbq	1ede	1fup	1han
1abr	1bhs	1cid	1ddt	1edg	1gad	1har
1ade	1bia	1ciy	1dea	1edh	1gai	1hav
1aep	1bmf	1cks	1def	1efn	1gal	1hbq
1aer	1bmt	1clc	1dek	1efu	1gar	1hce
1afb	1bnc	1cmb	1dhp	1eny	1gca	1hcn
1alo	1bnd	1cns	1dhr	1eri	1gcb	1hcp
1amm	1bp2	1cnt	1din	1erw	1gdo	1hcz
1amp	1bpl	1cof	1dja	1esc	1gdy	1hfh
1anv	1bri	1col	1dkz	1esf	1gen	1hge
1aoc	1bro	1cpc	1dlh	1esl	1ghr	1hgx
1aor	1buc	1cpo	1dnp	1etp	1gky	1hjr
1aoz	1bur	1cpq	1doi	1eur	1glc	1hlb
1apy	1bvp	1cpt	1dor	1ext	1gln	1hmy
1arb	1bw4	1crl	1dpe	1fba	1gnd	1hng
1ars	1byb	1csh	1dpg	1fbr	1gof	1hqa
1arv	1cau	1csm	1drw	1fc2	1got	1hrd
1ash	1ccr	1csn	1dsb	1fcd	1gp1	1hsl
1atl	1cdo	1ctn	1dsn	1fib	1gpb	1htm
1axn	1cdq	1ctt	1dsu	1fie	1gpc	1htp
1ayl	1cdy	1cur	1dts	1fil	1gph	1htt
1bam	1cea	1cus	1dup	1fim	1gpl	1huc
1bbh	1cel	1cvl	1dyn	1fjm	1gpm	1huw
1bbp	1cem	1cxs	1dyr	1fkj	1gpr	1hxn
1bbt	1ceo	1cyd	1eaf	1fkx	1grj	1hxp
1bcf	1cew	1cyg	1eal	1fnc	1grx	1i1b
1bco	1cfb	1cyv	1eca	1fnf	1gsa	1iae
1bdm	1cfr	1cyw	1ece	1frp	1gtq	1ice
1icn	1lba	1mka	1npo	1pii	1rci	1snc
1ign	1lbd	1mla	1oac	1pkm	1rec	1spg
1ihf	1lbi	1mld	1obp	1pkp	1reg	1sra
1ilk	1lbu	1m1s	1occ	1pls	1req	1sri
1ino	1lci	1mml	1oct	1pmi	1rfb	1std
1inp	1lck	1mol	1ofg	1pmy	1rgs	1stm
1iow	1lcl	1mpp	1omp	1pnk	1rhg	1svb
1irk	1lcp	1mrj	1onc	1poc	1rie	1svp
1irl	1ldm	1msa	1onr	1pot	1rnl	1svr
1isc	1leh	1msc	1ord	1pox	1rpa	1tag
1iso	1len	1msf	1oro		1rrg	1tal
1itg	1lfa	1msk	1osp	1pre	1rsy	1taq
1ivd	1lgr	1msp	1otg	1prs	1rtp	1tbr
1jap	1lis	1mty	1oun	1prt	1rva	1tca
tjcv	1lit	1mup	1ova	1psd	1sac	1tcr
1jon	1lki	1mut	1oxa	1ptv	1sat	1tfe
1jud	1lnh	1nal	1oxy	1ptx	1sbp	1tfr
1jvr	1lrv	1nar	1oyc	1pue	1sch	1tgx
1kaz	1ltd	1nba	1pbe	1put	1scm	1thj
1kcw	1lts	1ndh	1pbg	1pvc	1scu	1tht
1knb	1luc	1nfa	1pbn	1pvd	1sei	1thv
1kny	1lxa	1nfk	1pbw	1pxt	1ses	1thx
1kob	1lyl	1nfn	1pco	1pya	1sfe	1tii
1kpb	1mas	1nfp	1pdg	1qap	1sft	1tiv
1kpt	1mat	1nhk	1pdn	1qas	1slt	1tlk
1kuh	1maz	1nhp	1pea	1qba	1slu	1tml
1kve	1mda	1nif	1pfk	1qor	1sly	1tnr
1kxu	1mey	1nip	1pgs	1qpg	1smd	1tpg
1l48	1mhc	1nox	1phg	1rbu	1sme	1tpl
1lau	1mhl	1noy	1phr	1rcb	1smn	1trk
1tsp	1xjo	2bpa	2hts	2sil	4mt2	2aza
1ttb	1xnb	2btf	2hvm	2stv	4rhv	2bbv
1tul	1xrb	2cae	2ihl	2tbd	4sbv	2bgu
1tup	1xsm	2cas	2kau	2tct	4ts1	2blt
1tys	1xva	2cba	2lbp	2tgi	4xia	2bnh
1uby	1xyz	2cbp	2lfb	2tmd	5p21	2bop
1ucy	1ydr	2ccy	2mev	2tmv	5rub	2hft
1ulp	1yha	2cpl	2mnr	2tys	5tim	2hhm
1uxy	1ypp	2ctc	2mpr	2vil	6fab	2hmx
1vdc	1ytb	2cyp	2mta	3chy	7rsa	2hmz
1vhh	1ytw	2dkb	2nac	3cla	8abp	2hpd
1vhi	1yua	2dld	2olb	3cox	8acn	2hpe
1vhr	1znb	2dri	2omf	3dfr	8atc	2psp
1vid	1zqa	2ebn	2ora	3dni	8fab	2reb
1vin	2aaa	2end	2pcd	3fru	8tln	2rsi
1vls	2aakent	2eng	2pec	3geo	9pap	2rsp
1vmo	2abd	2er7	2pgd	3grs	9rnt	2sas
1vnc	2abh	2fal	2phy	3hhr	9wga	2scp
1vol	2abk	2fcr	2pia	3kin	1wba	3sdh
1vorm	2acq	2fd2	2pii	3min	1wdc	4aah
1vpt	2adm	2gmf	2pld	3pga	1whi	4bcl
1vsd	2ak3	2gsq	2pol	3pgm	1wht	4enl
1vsg	2amg	2gst	2por	3pmg	1xel	4fgf
1wad	2ayh	2hbg	2prk	3pte	1xik	4kbp

For the β-only proteins data set, 26 files were extracted with the criterion that the proteins do not contain α- helical structure; for the α-only proteins, 54 files were extracted with the criterion that they do neither contain Ba- or Bp-structured residues.

The α-bank corresponds to the following files:

1aep,1arv,1ash,1axn,1bbh,1bcf,1ccr,1cem,1cns,1cnt,1col,1cpq, 1eca,1etp,1hlb,1huw, 1ign,1ilk,1jvr,1kxu,1lis,1lki,1lrv,1maz,1mey, 1mls,1msf,1nfn,1oct,1pbw,1poc,1rci,1rfb,1rhg,1spg,1sra,1uby, 1vin,1vls,1xsm,2abd,2abk,2ccy,2cyp,2end,2fal,2hbg,2hmz,2lfb, 2tct,2tmv,3sdh,6fab,9wga.

The β-bank corresponds to the following files:

1cea,1cur,1fbr,1fnf,1hce,1i1b,1knb,1lcl,1msa,1msp,1nfa,1npo, 1pco,pdg,1ptx,1svp,1tgx,1tpg,1tul,1ulp,1vmo,1wba,1yha,2mpr, 2pii,4fgf.

ASA-Pex files

Each Pex file originates from the PDB file of a protein. Each line of the Pex corresponds to a residue in the order of the sequence, and each column is a parameter calculated in the 3D structure as described by Thomas (Thomas et al. 2001). In the ASA-Pex file, calculation of accessible surface areas (ASAs) of the whole residue (lateral chain and backbone) was achieved using the method of Shrake and Rupley (1973). In brief, the spherical surface of each atom is covered by a net of 642 points (the initial method used 92 points), and the points that lie within other expanded atoms are determined. The SERF algorithm where this method is implemented was used (Flower 1997).

Determination of the hydrophobic and hydrophilic accessible surfaces of residues

The method of Shrake and Rupley (1973) was also used to calculate the hydrophobic and hydrophilic ASAs. With this method, the

Petitioner Merck, Ex. 1035, p. 1415

hydrophobic and hydrophilic ASAs correspond to the sum of the surface of hydrophobic or hydrophilic atoms of the residue, respectively. Atoms are considered hydrophobic or hydrophilic depending on their transfer energy, as described (Brasseur 1991).

Median values of total, hydrophobic, and hydrophilic accessible surfaces were calculated using Excel software (Microsoft).

Determination of the secondary structure of the residues

In the PDB files, no protein has a complete description of secondary structures. We therefore established our definition of secondary structures based on the (phi/psi) values and on the occurrence of main chain H-bonds (O..H distance less than 3.5 Å), as described in Thomas et al (2001). Two different structures were attributed to the same residue according to the fact that that residue can be involved in two H-bonds, one on its NH side, the other on its CO side. Both secondary structures are listed in the Pex. In the present study, the NH secondary structure was considered.

Definition of the different secondary structures

Helices

Helical residues have a main chain H-bond and ϕ/ψ within a circle of 45° around the couple $\phi = -57°$ and $\psi = -47°$. The main chain H-bond has an O . . . H distance less than 3 Å. The helix is α (Ha) when the n and the n ± 4 residues are H-bound.

β-Structures

Residues are in β when the ϕ /ψ are within a circle of 90° around the $\phi = -129°$ and $\psi = 123°$. When two strands are H-bound, they are either antiparallel (Ba) or parallel (Bp) sheets. The sheets are parallel when the vectors between the Cα of the residues n and n + 1 of each strand draw an angle of −90° to +90°. They are antiparallel when the same vector angles are between 90° and 180° or −90° and −180° apart. B is for β-strands that are either not H-bound or are H-bound to a structure that is not a strand.

Random coils/turns

The ϕ/ψ values of the different turns are from Srinivasan and Rose (1995). The presence of an H-bond is not mandatory.

The ϕ/ψ of random coils span a large range of values, including the left helices that were not individualized in this study. Random coils also account for right helices and β-residues when the H-acceptor (the O = C residue, n + i) and the NH-donor (n) are too far away in the sequence for a helix (i > 6) or too close for a β-sheet (i < 3).

Molecular hydrophobicity potential (MHP) calculations

MHP is a three-dimensional plot of the hydrophobicity potential of a molecule created to visualize its amphipathy. The hydrophobicity of a molecule is calculated using its partition coefficient between water and octanol.

We postulate that the hydrophobicity induced by an atom i and measured at a point M of space decreases exponentially with the distance between this point M and the surface of an atom i according to the equation (Brasseur 1991):

$$MHP_M = \sum_{l=1}^{N} E_{tri} \exp (r_i - d_i)$$

where N is all atoms of the molecule, E_{tri} is the transfer energy of atom i, r_i is the radius of the atom i, and d_i is the distance between atom i and the point M. E_{tri} is the energy required to transfer an atom i from a hydrophobic to a hydrophilic medium. Atomic E_{tri} values were calculated from the molecular transfer energies compiled by Tanford (1973), assuming that molecular E_{tr} are the sum of their atomic E_{tr}. Atomic E_{tr} values were derived for seven different atom types (Brasseur 1991).

The hydrophobic and hydrophilic isopotential surfaces were calculated by a cross-sectional computational method. A 1-Å meshgrid plane was set to sweep across the molecule by steps of 1 Å. At each step, the sum of the hydrophobicity and hydrophilicity values at all grid nodes was calculated. The hydrophobic and hydrophilic MHP surfaces were then drawn by joining the isopotential values.

All calculations were performed on Pentium III processors, using Z-TAMMO and Z-PEX software. Molecular graphs were drawn using WinMGM (Ab Initio).

Acknowledgments

R.B. and L.L. are Research Director and Research Associate, respectively, at the Belgian Funds for Scientific Research (FNRS). A.T. is Research Director at the French National Institute for Medical Research (INSERM). This work was supported by Region Wallone (Grants 114.830 and 14.540 to R.B.) and the Belgian Fonds de la Recherche Scientifique Médicale (grant 3.4568.03 to L.L.). We thank Dr. D.R. Flower for kindly providing the SERF algorithm.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked " advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

Brasseur, R. 1991. Differentiation of lipid-associating helices by use of three-dimensional molecular hydrophobicity potential calculations. *J. Biol. Chem.* **266:** 16120–16127.

Chothia, C. 1976. The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* **105:** 1–12.

Chou, K.C., Zhang, C.T., and Maggiora, G.M. 1997. Disposition of amphiphilic helices in heteropolar environments. *Proteins* **28:** 99–108.

Creighton, T. 1993. Chemical properties of polypeptides. In *Chemical properties of polypeptides* (ed. T. Creighton), pp. 1–46. W.H. Freeman and Company, New York.

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

Flower, D.R. 1997. SERF: A program for accessible surface area calculations. *J. Mol. Graph. Model.* **15:** 238–244.

Jones, S. and Thornton, J.M. 1997. Analysis of protein–protein interaction sites using surface patches. *J. Mol. Biol.* **272:** 121–132.

Li, L., Darden, T.A., Bartolotti, L., Kominos, D., and Pedersen, L.G. 1999. An atomic model for the pleated β-sheet structure of Aβ amyloid protofilaments. *Biophys. J.* **76:** 2871–2878.

Li, X. and Pan, X.M. 2001. New method for accurate prediction of solvent accessibility from protein sequence. *Proteins* **42:** 1–5.

Liu, W. and Chou, K.C. 1999. Prediction of protein secondary structure content. *Protein Eng.* **12:** 1041–1050.

Naderi-Manesh, H., Sadeghi, M., Arab, S., and Moosavi Movahedi, A.A. 2001. Prediction of protein surface accessibility with information theory. *Proteins* **42:** 452–459.

Petitioner Merck, Ex. 1035, p. 1416

Rost, B. 1996. PHD: Predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol.* **266:** 525–539.
Rost, B. and Sander, C. 1994. Conservation and prediction of solvent accessibility in protein families. *Proteins* **20:** 216–226.
Samanta, U., Bahadur, R.P., and Chakrabarti, P. 2002. Quantifying the accessible surface area of protein residues in their local environment. *Protein Eng.* **15:** 659–667.
Schladitz, C., Vieira, E.P., Hermel, H., and Mohwald, H. 1999. Amyloid-β-sheet formation at the air–water interface. *Biophys. J.* **77:** 3305–3310.
Shrake, A. and Rupley, J.A. 1973. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.* **79:** 351–371.
Srinivasan, R. and Rose, G.D. 1995. LINUS: A hierarchic procedure to predict the fold of a protein. *Proteins* **22:** 81–99.
Tanford, C. 1973. In *The hydrophobic effect: Formation of micelles and biological membranes* (ed. C. Tanford), pp. 5–20. Wiley, New York.
Thomas, A., Bouffioux, O., Geeurickx, D., and Brasseur, R. 2001. Pex, analytical tools for PDB files. I. GF-Pex: Basic file to describe a protein. *Proteins* **43:** 28–36.
Thomas, A., Meurisse, R., Charloteaux, B., and Brasseur, R. 2002a. Aromatic side-chain interactions in proteins. I. Main structural features. *Proteins* **48:** 628–634.
Thomas, A., Meurisse, R., and Brasseur, R. 2002b. Aromatic side-chain interactions in proteins. II. Near- and far-sequence Phe-X pairs. *Proteins* **48:** 635–644.
Yuan, Z., Burrage, K., and Mattick, J.S. 2002. Prediction of protein solvent accessibility using support vector machines. *Proteins* **48:** 566–570.

Petitioner Merck, Ex. 1035, p. 1417