Protein Science (1994), *3*:717–729. Cambridge University Press. Printed in the USA.

A role for surface hydrophobicity in protein–protein recognition

L. YOUNG,[1] R.L. JERNIGAN,[1] AND D.G. COVELL[2]

[1] Laboratory of Mathematical Biology, Division of Cancer Biology Diagnosis and Centers, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892
[2] Biomedical Supercomputing Laboratory, PRI/Dyncorp, Frederick Cancer Research and Development Center, National Cancer Institute, Frederick, Maryland 21702

(RECEIVED December 2, 1993; ACCEPTED March 14, 1994)

Abstract

The role of hydrophobicity as a determinant of protein–protein interactions is examined. Surfaces of apo-protein targets comprising 9 classes of enzymes, 7 antibody fragments, hirudin, growth hormone, and retinol-binding protein, and their associated ligands with available X-ray structures for their complexed forms, are scanned to determine clusters of surface-accessible amino acids. Clusters of surface residues are ranked on the basis of the hydrophobicity of their constituent amino acids. The results indicate that the location of the co-crystallized ligand is commonly found to correspond with one of the strongest hydrophobic clusters on the surface of the target molecule. In 25 of 38 cases, the correspondence is exact, with the position of the most hydrophobic cluster coinciding with more than one-third of the surface buried by the bound ligand. The remaining 13 cases demonstrate this correspondence within the top 6 hydrophobic clusters. These results suggest that surface hydrophobicity can be used to identify regions of a protein's surface most likely to interact with a binding ligand. This fast and simple procedure may be useful for identifying small sets of well-defined loci for possible ligand attachment.

Keywords: binding sites; CD4; HIV; HLA; hydrophobicity; residue–residue interactions

Modern studies of biological phenomena attempt to establish the relationship between structure and function at the molecular level. One widely studied mechanism underlying most biological processes is the molecular recognition exhibited by the association of a substrate and its target enzyme. These associations involve interactions between accessible portions of each molecule's surface and are thought to be determined largely by the details of geometric and chemical complementarity (Chothia & Janin, 1975; Dill, 1990; Sharp et al., 1991a, 1991b).

Discovering the important details underlying geometric and chemical complementarity has been a long-standing research goal that has stimulated a wide variety of approaches. Regions of high surface complementarity (Shoichet et al., 1992) have been used for choosing the best molecular geometry for docking (Kuntz et al., 1982). Nicholls et al. (1991) propose surface curvature as a useful tool for locating binding sites. Regions with high densities of hydrogen bond sites (Danziger & Dean, 1989a, 1989b) and with favorable electrostatic properties (Goodford, 1985) have been used to optimize chemical complementarity as well as to search selected surfaces for possible binding sites. Quantum chemical and statistical mechanical studies using the algorithmic tools of molecular dynamics simulations (Brooks et al., 1988), semi-empirical methods (Rullmann & van Duijen, 1990), and density functional methods (Bethe, 1993; Johnson et al., 1993) have also been applied within the framework of theoretical chemistry to understand better the essential characteristics of molecular associations (Janin & Chothia, 1990).

Studies on the related question of protein unfolding and hydrocarbon solubilities in water (Tanford, 1980; Privalov & Gill, 1988) suggest that the strengths of amino acid associations originate from interactions between preferred hydrophobic atoms (Dill, 1990). The close-packed arrangement of interacting protein molecules has been suggested as additional thermodynamic evidence (Cherfils et al., 1991; Varadarajan et al., 1992; Kelley & O'Connell, 1993) that the basis of these interactions resides partly in the hydrophobic effect (Kauzmann, 1959). Support for this claim is found in the strong correlations between buried surface area and measured binding strengths of many protein–protein complexes (Horton & Lewis, 1992). The recent review of Cherfils et al. (1991), however, notes that when compared with the rest of an enzyme's surface, regions involved in interactions with other molecules are neither more hydrophobic nor enriched in groups bearing an electric charge.

The present study reexamines the role of hydrophobicity as a determinant of the preferred sites of molecular associations and attempts to quantify its importance in identifying substrate

Reprint requests to: D.G. Covell, PRI/DynCorp, NCI–Frederick Cancer Research and Development Center, P.O. Box B, Frederick, Maryland 21702-1201; e-mail: covell@ncifcrf.gov.

attachment sites. The analysis is restricted to proteins complexed with proteins, peptides, and peptide-like fragments currently available in the Brookhaven database (Bernstein et al., 1977; Abola et al., 1987). The surface of each apo-form of these molecules is scanned to determine clusters of surface-accessible amino acids. Each cluster is scored according to the hydrophobicity of its constituent amino acids. The positions of clusters with the best hydrophobicity scores are compared to the portion of surface buried by the ligand as it appears in the complexed form. The results indicate that, for the set of molecules tested, wide variations are found in the hydrophobicity scores for clusters on each surface. Clusters with the best hydrophobicity scores are usually found to include the attachment site for the complexed molecule. This strong correlation between the attachment site and the most hydrophobic clusters supports the concept that protein–protein associations usually involve the more hydrophobic portions on the surface of a molecule, notwithstanding the fact that the detailed boundary of the site and the precise orientation with respect to the ligand can depend on further properties beyond hydrophobicity. The analysis is simple, does not require calculation of molecular surfaces, and is extremely fast to implement on most workstations.

Results

Analysis of surface hydrophobicity

The molecules included in this analysis are listed in Table 1 and are described further in the Materials and methods section at the end of the paper. Each molecule in a complex has been analyzed in the absence of its binding partner. The results are grouped according to molecular size such that Tables 2 and 3 and Tables 4 and 5 present the results for the larger and smaller molecules in the complex, respectively.[3]

The results for the target molecules indicate that the number of clusters identified on their surface ranges from 1,530 to 3,187 (Table 2, column 2). The number of amino acids in a cluster ranges from 3 to 15. Each cluster is ranked according to its total hydrophobicity, with the most hydrophobic cluster at the top of the list. Column 3 of Table 2 identifies the highest ranked cluster with at least a 30% overlap with the surface buried by the bound substrate. In other words, the value in this column indicates the ranking of the most hydrophobic cluster with greater than 30% overlap with the position of the known ligand. The clear trend of these results indicates a high correspondence between the region of greatest hydrophobicity and the binding interface with the complexed substrate. In 12 cases the correspondence is exact, with the most hydrophobic cluster having an average of 65% overlap with the surface buried by the ligand. Two ranked clusters are shown for 4TPI because the Val-Val fragment in the crystallographic structure was found to correspond to the position of the highest ranked hydrophobic cluster. Two hydrophobic clusters that overlap the binding site for 2CPK are also found at rankings 1 and 5. These clusters correspond to the 2 ends of 2CPK's large elongated binding site, which has hydrophobic interactions with the peptide inhibitor PKI(5-24) (Knighton et al., 1991).

This analysis reveals that the active site of the 9 distinct types of enzyme complexes[4] is identified correctly for at least 1 member of each enzyme family within the first 3 listed sites. The correspondence for antibody fragments finds this ranking within the top 6 clusters. The remaining 3 molecules not included in these 2 groups have rankings of their most favorable cluster within the top 4 positions. These results suggest that simple measures of hydrophobicity can be used to identify a few regions on a molecule's surface, one or more of which are found to share a portion of the surface involved in binding with a known ligand. An example of this correspondence is shown in Figure 1 for the enzymes 4SGB and 1TGS. The portion of the enzyme surface associated with the 2 highest ranked hydrophobic clusters, in the case of 4SGB, and the most hydrophobic cluster, in the case of 1TGS, is colored spectrally according to the hydrophobicity of the amino acids composing these clusters. Residues with the greatest hydrophobicity are indicated by red and those with the least by purple. The strongest hydrophobic cluster for 4SGB consists of the amino acids Ala 164, Thr 168, Val 169, Arg 182, Gly 215, Gly 216, Ser 217, Gly 224, Thr 225, Thr 226, and Phe 227. This cluster includes 9 of the 21 amino acids found in the interface. When viewed from the orientation shown in this figure, these residues compose the lower portion of the valley into which the ligand binds. The second strongest cluster consists of residues Ser 48, Thr 51, Tyr 52, Ser 88, Gly 89, Val 106, Arg 107, Tyr 108, Tyr 238, and Gly 239. These residues form the upper portion of the valley into which the ligand binds and include 10 of the remaining 12 amino acids in the interface. The strong hydrophobic character of these clusters is the result of contributions from valine, phenylalanine, alanine, tyrosine, and to a lesser extent the 5 glycines. Residues 31–42 of the inhibitor ligand bound to 4SGB form the closest interactions with the target enzyme residues and are shown as a ball-and-stick model in Figure 1A. This stereo view clearly indicates close interaction between this portion of the inhibitor and the set of amino acids forming these 2 clusters. A similar picture is seen for the enzyme 1TGS (Fig. 1B). The best ranked hydrophobic cluster having the greatest correspondence with the interface was second on the list of 1,766 possible clusters. This cluster consists of residues Ala 56, His 57, Tyr 59, Ile 89, Val 90, His 91, Pro 92, Ser 93, Tyr 84, and Ser 96. These residues are included as part of the amino acids that form a pocket for interaction with segment 11–22 of the trypsin inhibitor. The view shown in this figure indicates that the hydrophobic residues of this cluster appear to wrap around the central portion of this segment of the inhibitor. In all cases studied, the correspondence between strongly hydrophobic clusters and the bound substrate is clearly evident.

Analysis of buried surfaces

Accessible surface areas are used only to compare the results of this calculation with the observed X-ray data. Actual and calculated surfaces are defined as the accessible surface area of the protein buried by the ligand (A_{actual}) and that associated with any hydrophobic cluster ($A_{cluster}$), respectively. Comparison of these 2 surfaces determines how well each cluster in the ranked list corresponds to the actual binding site. Any measure of over-

[3] This designation serves only to separate the larger molecules (either an enzyme or antibody) from the smaller molecules (either an inhibitor or antigen).

[4] Trypsin serine protease, elastase serine protease, chymotrypsin serine protease, subtilase serine protease, eukaryotic aspartyl protease, zinc protease hydrolase, transferase, acid proteinase, and carboxypeptidase.

Table 1. *Molecules included in the analysis*

PDB	Target: Name	Target: No. amino acids	Ligand: Name	Ligand: No. amino acids
	A. Enzyme complexes			
	Trypsin serine protease			
1TPA	Anhydro-trypsin (E.C. 3.4.21.4)	223	Pancreatic trypsin inhibitor	58
2PTC	β-Trypsin (E.C. 3.4.21.4)	223	Pancreatic trypsin inhibitor	58
1TGS	Trypsinogen (E.C. 3.4.21.4)	225	Porcine pancreatic secretory trypsin inhibitor	56
2TGP	Trypsinogen (E.C. 3.4.21.4)	223	Pancreatic trypsin inhibitor	58
4TPI	Trypsinogen (E.C. 3.4.21.4)	223	[Arg 15] pancreatic trypsin inhibitor	58
4TPI	Trypsinogen (E.C. 3.4.21.4)	223	Val-Val	2
4SGB	Serine protease B (E.C. 3.4.21.4)	185	Potato inhibitor	51
2KAI	Kallikrein A (E.C. 3.4.21.4)	232	Bovine pancreatic trypsin inhibitor	56
	Elastase serine protease			
1HNE	Human neutrophil elastase (E.C. 3.4.21.-)	218	Methoxysuccinyl-Ala-Ala-Pro-Ala chloromethyl ketone	5
2EST	Elastase (E.C. 3.4.21.11)	240	TFA-Lys-Ala_ANI	4
	Chymotrypsin serine protease			
1CHO	α-Chymotrypsin (E.C. 3.4.21.1)	237	Turkey ovomucoid third domain (OMTKY3)	53
	Subtilase serine protease			
2SEC	Subtilase Carlsberg (E.C. 3.4.21.14)	274	Genetically engineered *N*-acetyl eglin-C	64
2SNI	Subtilisin novo (E.C. 3.4.21.14)	275	Chymotrypsin inhibitor 2	64
1TEC	Thermitase (E.C. 3.4.21.14)	279	Eglin-C	63
	Eukaryotic aspartyl protease			
2ER9	Endothelial aspartic proteinase (E.C. 3.4.23.6)	328	L363,564	6
5APR	Acid proteinase (*Rhizopus* pepsin) (E.C. 3.4.23.6)	330	Pepstatin-like renin inhibitor	6
	Zinc protease hydrolase			
6TMN	Thermolysin (E.C. 3.4.24.4)	316	Cbz-Gly(P)-(O)-Leu-Leu(ZG(P)(O)LL)	3
1TLP	Thermolysin (E.C. 3.4.24.4)	316	Phosphoramidon	3
	Transferase			
2CPK	Cyclic AMP-dependent protein kinase (E.C. 2.7.1.37)	336	Peptide inhibitor PKI(5-24)	20
	Acid proteinase			
4HVP	HIV-1 protease (E.C. 3.4.23.-)	198	*N*-acetyl-Thr-Ile-Nle-psi[CH_2-NH]-Nle-Gln-Arg-amide (MVT101)	6
	Carboxypeptidase			
4CPA	Carboxypeptidase A (E.C. 3.4.17.1)	307	Potato carboxypeptidase A inhibitor	37
	Protein: Name	Protein: No. amino acids	Ligand: Name	Ligand: No. amino acids
	B. Protein complexes			
	Antibody			
1FDL	IgG1 Fab fragment (D1.3 κ)	432	Lysozyme (E.C. 3.2.1.17)	129
2HFL	IgG1 Fab fragment (HyHEL-5)	425	Lysozyme (E.C. 3.2.1.17)	129
3HFM	IgG1 Fab fragment (HyHEL-10)	429	Lysozyme (E.C. 3.2.1.17)	129
	Fab′ fragment			
1HIM	IgG2a fragment	431	Synthetic decamer peptide	8
2IGF	IgG1 Fab′ fragment	440	Residues 69-87 of myohemerythrin	19
1NCB	Fab′ complex	435	Neuraminidase (E.C. 3.2.1.18)	389
	FC fragment			
1FC2	IgG FC fragment	206	Fragment B of protein A	43
	Hirudin			
1HTC	Hirudin variant 2-lysine	291	α-Thrombin	65
	Growth hormone			
2HHR	Portion of growth hormone	379	Portion of extracellular receptor domain	194
	Retinol binding protein			
1RBP	Retinol binding protein	174	Retinol	1

Table 2. *Hydrophobic cluster rankings for target molecules*

PDB	No. points	Rank	% Rank
1TPA	1,713	3	0.18
2PTC	1,704	3	0.18
1TGS	1,766	2	0.12
2TGP	1,758	3	0.17
4TPI	1,741	3	0.17
(4TPI)	1,741	1	0.06
4SGB	1,530	1	0.07
2KAI	1,794	2	0.11
1HNE	1,769	2	0.12
2EST	1,871	1	0.05
1CHO	1,752	2	0.11
2SEC	1,769	4	0.23
2SNI	1,805	4	0.22
1TEC	1,819	3	0.16
2ER9	2,346	3	0.13
5APR	2,288	3	0.13
6TMN	2,167	1	0.05
1TLP	2,225	1	0.04
2CPK	2,435	1	0.04
4HVP	1,715	1	0.06
4CPA	2,095	1	0.05
1FDL	3,119	6	0.19
2HFL	3,107	4	0.13
3HFM	3,077	1	0.03
1HIM	3,187	2	0.06
2IGF	3,168	3	0.09
1NCB	3,038	2	0.07
1FC2	2,079	1	0.05
1HTC	2,230	1	0.04
2HHR	3,021	4	0.17
1RBP	1,648	1	0.06
Average		2.3	0.11
SD		1.3	0.06

lap between these regions depends not only on the extent of their coincidence but also on their total size; overpredicting the size of a patch will falsely enhance coincidence with the surface buried by the known ligand. An ideal measure would distinguish surface regions of exactly the same size in perfect register from regions of differing size and position. In order to retain information about size and position between differing surfaces, 2 measures (M_1 and M_2) of percentage overlaps are used:

$$M_1 = \frac{A_{cluster} \cap A_{actual}}{A_{actual}} \times 100 \quad (1)$$

$$M_2 = \frac{A_{cluster} \cap A_{actual}}{A_{cluster}} \times 100. \quad (2)$$

The larger the measure M_1, the better is the coverage of the actual binding site by the hydrophobic cluster. These measures of percentage overlap for the 2 surfaces—actual and calculated—taken together indicate the relative sizes of surface hydrophobic clusters and actual binding sites. High values for both M_1 and M_2 identify similar-sized overlapping areas. A high percentage for M_1, accompanied by a lower percentage for M_2, corresponds to a hydrophobic cluster larger than the actual binding site. Conversely, low percentages for M_1 with higher percentages for M_2 imply the actual binding site is larger than the calculated hydrophobic cluster. The ideal is 100% for both M_1 and M_2, corresponding to the situation where the hydrophobic cluster corresponds precisely with the same protein surface buried by the known ligand.

Table 3 lists the measures of overlap, M_1 and M_2, for the 4 top-ranked clusters of the target molecules. In the discussion to follow, these hydrophobic clusters will be referred to as the calculated sites for each protein. A hydrophobic cluster with M_1 >30% appears within the 4 top-ranked hydrophobic clusters for each target molecule, with 1 exception. The exception is for the anti-lysozyme Fab′ fragment 1FDL bound to lysozyme, where this cluster ranking is sixth in the list. The values of M_1 range from 30.5 to 99.7% for the highest ranked clusters that have greater than 30% overlap with the actual binding site. In most instances, at least 2 of the 4 top-ranked hydrophobic clusters overlap the actual binding site (i.e., non-zero values for M_1 and M_2). Cases exist where only 1 cluster in the top 4 clusters corresponds with the actual binding site (2KAI, 2EST, and 3HFM). The group average of M_1 for each of the top 4 clusters in all cases studied was 24.3 ± 29.9%. This suggests that, on average, some overlap between the predicted and actual binding site is found within the 4 top-ranked hydrophobic clusters for all molecules tested. The corresponding values of M_2 range from 5.3 to 84.8%, with a group average for each of the 4 top rankings of 14.9 ± 20.1%. These results indicate that the calculated site tends to cover a larger surface than the actual site. The extent of this additional surface represents, however, only a small fraction of the protein's total accessible surface. The size of the actual binding sites as a percentage of the total surface area of the target molecule is listed in the last column of Table 3 (labeled $\%A_{actual}$). An average of 5.3% of the target molecule's total surface area is buried in the actual binding interface, the largest being 11.4% and the smallest 1.6%. The calculated sites have their average percentage of total surface area buried by each of the top 4 clusters as 7.5 ± 2.6%. This slightly larger coverage of the surface by the predicted site, when compared to the actual site, contributes to the lower values of M_2 found in this table.

Analysis of ligands

The results obtained from analysis of the ligands to these molecules are summarized in Table 4. These results reflect those found for the apo-forms of the target molecules and indicate a stronger correspondence between the most hydrophobic cluster and the binding interface. The analysis finds that the strongest hydrophobic cluster corresponds to the actual binding interface in 13 of the 18 cases examined.[5] The average ranking of those clusters having at least 30% overlap with the actual binding site was 1.7. This corresponds to an average percentile ranking of 0.18% for the strongest hydrophobic cluster that exhibits over 30% overlap with the actual interface. This corresponds to nearly a 3-order-of-magnitude reduction in the total number of surface clusters to be considered before a portion of the actual binding interface is located. The lower rankings for LYZ-2HFL

[5] This analysis was not performed on the smallest peptide inhibitors because the distances used to assign residues to clusters usually include most of the structure and thus limit comparisons between clusters.

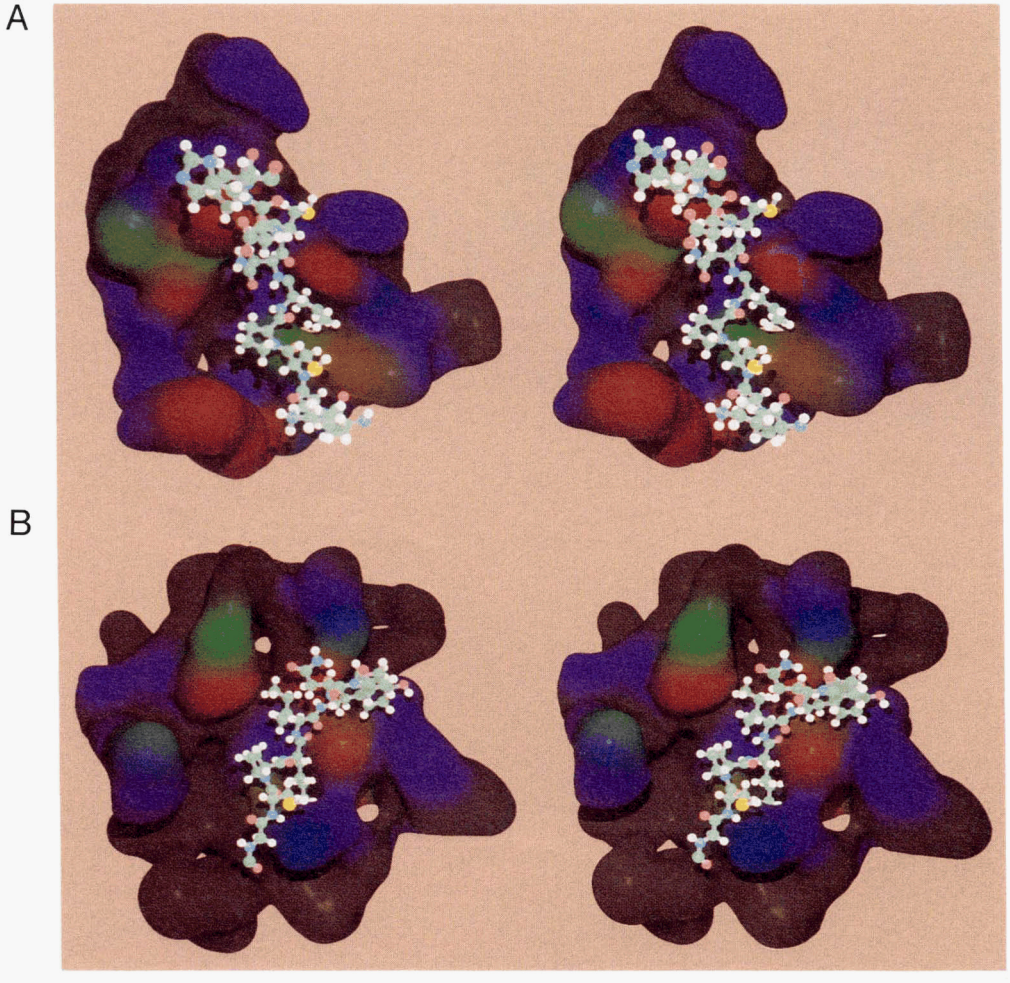

Fig. 1. Stereo views of (**A**) serine proteinase B (4SGB) and residues 34–42 of the 51-residue potato inhibitor (ball-and-stick representation) and (**B**) trypsinogen (1TGS) and residues 11–22 of the 56-residue porcine pancreatic secretory trypsin inhibitor. Enzyme surfaces are colored spectrally, with red and purple representing the most and least hydrophobic amino acids in a cluster, respectively; brown residues are not included in a cluster. The non-brown-colored regions of 4SGB comprise the 2 most hydrophobic clusters found on its surface. (These figures were generated using Rayshade as modified by George McGregor, Program Resources Incorporated.)

(rank 4) and LYZ-3HFM (rank 5) are consistent with the observations that the anti-lysozyme antibodies analyzed here bind to 3 different portions of the lysozyme surface.

The fractional overlap of the actual binding surface and that associated with the clusters having the best hydrophobicity scores are listed in Table 5. On average the values for M_1 and M_2 were slightly higher than those found in Table 3. The group average for M_1 of 35.8 ± 26.4% and M_2 of 30.1 ± 21.4% are 11.5% and 15.2% greater than the corresponding averages for the set of target molecules. The average value of M_1 for each of the 4 most hydrophobic clusters having >30% overlap ranged from 46.5 to 24.3%, whereas the corresponding averages of M_2 ranged from 35.6 to 21.4%. A group average of 20.3 ± 14.1% of the molecule's accessible surface is involved in binding (Table 5, column labeled $\%A_{actual}$). Similar values are found for the percentage of the molecule's surface associated with the 4 top-ranked hydrophobic clusters. Values of $\%A_{cluster}$ range from 21.4 to 23.4% of the ligand's surface with a group average for the 4 top-ranked hydrophobic clusters of 22.2 ± 8.9%. These percentages are approximately 3–4 times larger than those found for the apo-forms of the target molecules listed in Table 3 and reflect the smaller size of the ligand when compared to the target molecule.

Discussion

General

The results of this analysis indicate that simple measures of hydrophobicity applied to a C^{α} protein backbone model can be used to identify portions of a molecule's surface that may be involved in binding. For all complexes analyzed, the interface for binding corresponds to surface regions of strong hydrophobicity. The analysis is based only on knowledge of the target structure; no information about the substrate is required. The success of this procedure lends additional support to the notion that surface hydrophobic interactions participate strongly in molecular

Table 3. *Measures of overlap between surfaces buried by the 4 top-ranked hydrophobic clusters and actual binding interface for target molecules*

	M_1% (top 4)				M_2% (top 4)				%$A_{cluster}$ (top 4)				
PDB	1	2	3	4	1	2	3	4	1	2	3	4	%A_{actual}
1TPA	1.1	0.0	55.7	23.5	0.7	0.0	32.4	14.9	11.0	8.7	10.4	11.2	7.3
2PTC	0.0	0.0	34.9	11.9	0.0	0.0	19.2	6.7	10.8	8.2	13.1	12.4	7.2
1TGS	0.0	30.5	21.1	0.0	0.0	21.1	17.0	0.0	9.8	10.4	10.2	8.7	7.8
2TGP	0.0	0.0	48.1	16.4	0.0	0.0	27.1	17.1	8.3	8.5	12.1	6.8	6.7
4TPI	0.0	0.0	70.9	0.0	0.0	0.0	84.8	0.0	8.7	10.4	7.6	7.8	6.3
(4TPI)	54.6	27.7	1.5	0.9	3.7	5.3	0.5	0.2	8.7	10.4	7.6	7.8	1.6
4SGB	38.6	0.0	0.0	69.5	25.5	0.0	0.0	51.9	12.4	11.0	10.8	10.6	8.5
2KAI	0.0	63.3	0.0	0.0	0.0	44.5	0.0	0.0	9.2	8.8	8.9	7.8	6.2
1HNE	0.0	87.3	0.0	0.6	0.0	22.4	0.0	0.2	8.4	8.4	6.4	8.6	2.6
2EST	48.8	0.0	0.0	0.0	6.8	0.0	0.0	0.0	18.4	12.9	9.6	9.8	2.1
1CHO	0.0	60.1	0.0	37.9	0.0	34.7	0.0	28.9	7.6	10.3	8.1	8.3	5.9
2SEC	0.0	0.0	70.2	13.4	0.0	0.0	62.7	15.0	8.5	9.4	6.9	5.8	5.7
2SNI	0.0	17.8	8.0	31.5	0.0	17.6	26.3	46.9	7.0	7.5	5.7	8.1	7.2
1TEC	0.0	19.7	68.8	53.7	0.0	16.9	61.2	80.8	6.6	7.9	6.3	7.2	6.4
2ER9	0.0	23.0	43.3	26.2	0.0	12.4	24.1	17.8	5.1	9.6	5.8	7.7	3.2
6APR	0.0	65.9	0.0	44.6	0.0	34.9	0.0	33.1	6.4	7.9	8.9	5.6	3.3
6TMN	97.3	4.3	0.0	10.6	29.3	1.7	0.0	5.2	8.9	7.3	5.1	6.0	2.7
1TLP	99.7	88.4	84.1	0.0	40.6	27.1	29.5	0.0	6.7	9.5	8.5	8.1	2.7
2CPK	48.1	10.1	68.4	0.0	35.4	10.6	74.4	0.0	7.7	6.4	5.8	5.6	6.7
4HVP	52.4	31.4	1.4	15.9	16.7	12.7	1.1	11.8	9.8	10.8	7.8	9.1	3.1
4CPA	72.2	78.5	0.0	0.0	66.6	67.8	0.0	0.0	5.2	5.5	5.6	5.7	4.7
1FDL	0.0	0.0	0.0	9.2	0.0	0.0	0.0	5.7	6.2	5.1	3.8	4.9	3.2
2HFL	2.8	0.0	1.1	33.8	3.0	0.0	1.3	26.2	3.9	4.2	3.6	5.3	4.1
3HFM	70.4	0.0	0.0	0.0	52.2	0.0	0.0	0.0	5.2	6.5	5.7	5.4	3.8
1HIM	37.1	98.1	0.0	0.5	15.9	28.1	0.0	0.2	5.5	2.3	2.6	4.5	7.9
2IGF	28.9	0.0	92.8	39.8	23.8	0.0	33.2	16.6	5.5	5.5	2.2	3.1	6.1
1NCB	0.0	42.2	0.0	31.3	0.0	33.4	0.0	29.8	4.7	4.9	6.6	4.8	6.1
1FC2	95.3	19.9	0.0	55.0	48.3	11.2	0.0	36.2	5.8	10.7	7.0	8.2	11.4
1HTC	78.2	27.1	5.5	30.3	58.6	16.3	3.0	19.3	3.8	6.1	5.4	4.1	6.1
2HHR	0.0	0.0	1.5	31.1	0.0	0.0	0.3	13.3	5.3	6.1	6.2	5.8	4.2
1RBP	50.8	49.5	0.0	0.0	7.9	12.4	0.0	0.0	1.2	4.8	10.4	9.4	4.5
Average	28.4	27.4	22.2	19.5	14.7	14.4	16.1	16.0	7.6	8.0	7.0	7.0	5.3
SD	35.2	31.5	31.5	20.2	20.6	16.6	24.8	19.5	3.1	2.5	2.6	2.1	2.2
Group mean		24.3				14.9				7.5			
Group SD		29.9				20.1				2.6			

recognition. Furthermore, the present success provides evidence that peptides interact with similar strengths, whether intramolecularly in the interior of a folded protein or intermolecularly on its surface. In both cases, stabilization appears to result from burial of the most hydrophobic residues.

The argument arises that our results are simply a function of geometry such that, in order to define the most favorable attachment site, one need only find clusters with a large number of residues, e.g., concave sites on the protein. However, even though the most favorable clusters often have a larger number of protein residues than the majority of the clusters, selecting targets on the basis of the number of residues alone yields several clusters with the same number of residues. Table 6 compares the cluster rankings of the method explored in the present work with those given by ranking clusters according to largest number of residues in a cluster. In a majority of cases, ranking by the greatest number of residues, without weighting by some measure of energy or hydrophobicity, identifies a greater number of high-scoring clusters than found using the procedure described here. The present approach of combining a lattice model with calculation of hydrophobicities to identify important surface regions includes such geometric effects implicitly, permits significantly better discrimination, and avoids cluster degeneracy (i.e., several clusters with the same rank).

The statistics for the residues involved in the binding interface of each complex are summarized in Table 7. The amino acid composition for the proteins studied here is given in this table, together with their ratio to the frequencies of occurrence found in globular proteins (Table 1-1 of Creighton, 1984) listed in parentheses. No large differences are observed between the frequencies in these proteins and the more global frequencies. The most extreme cases differing from globular proteins include a lower than expected occurrence of glutamic acid and a higher than expected occurrence of tryptophan and serine. These differences are not thought to indicate any bias in the selected proteins distinct from that manifested in globular proteins in

Table 4. *Hydrophobicity cluster rankings for substrate molecules*

PDB	No. points	Rank	% Rank
1TPA	769	2	0.18
2PTC	783	1	0.12
1TGS	847	1	0.12
2TGP	789	1	0.13
4TPI	772	1	0.13
4SGB	800	1	0.13
2KAI	788	4	0.13
1CHO	792	1	0.13
2SEC	868	1	0.12
2SNI	902	1	0.12
1TEC	851	1	0.12
2CPK	581	1	0.17
4CPA	634	1	0.15
LYZ-1FDL	1,277	1	0.08
LYZ-2HFL	1,277	4	0.31
LYZ-3HFM	1,277	5	0.39
1NCB	2,350	2	0.09
2HHR	2,044	1	0.05
Average		1.7	0.18
SD		1.4	0.11

general. The high occurrence of serine may, however, be ascribable to the 13 serine proteases in the list of enzymes studied.

The occurrences of amino acids separately, at the binding interface for the ligand and for the target molecules are listed in columns 3 and 4 of Table 7, respectively.[6] The numbers in parentheses represent the ratio of each amino acid's observed to average frequency of occurrence (Creighton, 1984). Both ligand and target molecules have cysteine and tryptophan in their interfaces at frequencies at least 50% above their global averages. In addition, methionine, leucine, valine, alanine, glutamine, glutamic acid, and lysine appear between 70 and 20% less frequent than their averages for both target and ligand molecules. A portion of the high appearance of cysteine is the result of having pancreatic trypsin inhibitor in 7 of the cases studied. A clear explanation has not been found for the higher occurrence of glycine or any of the amino acid types that appear with lower frequencies. Observation of each molecule separately also finds the appearance of tyrosine, asparagine, and proline in the ligand molecules and glycine, threonine, serine, and histidine in the target molecules to be greater than 50% above their average values. These differences are more likely to reflect the limited sample size for the proteins studied (N(ligand) = 328, N(target) = 960) rather than important details about the amino acid composition at the binding interface. This observation, taken together with the earlier discussion about geometry, indicates that neither amino acid composition nor shape alone is a strong indicator of an attachment site. The combination of the two features appears to be a powerful tool for identifying a binding interface.

Cases exist where highly ranked hydrophobic clusters do not have any overlap with the bound ligand (cf. Tables 3, 5). This

[6] Amino acids with their C^{α} positions within 6.1 Å of the bound ligand are included in this set.

Table 5. *Measures of overlap between surfaces buried by the 4 top-ranked hydrophobic clusters and actual binding interface for substrate molecules*

	M_1% (top 4)				M_2% (top 4)				%$A_{cluster}$ (top 4)				
PDB	1	2	3	4	1	2	3	4	1	2	3	4	%A_{actual}
1TPA	15.5	46.8	36.9	0.0	11.3	47.2	33.8	0.0	26.4	19.9	21.1	22.1	20.2
2PTC	46.3	22.2	63.9	18.7	40.2	19.8	60.3	21.4	23.7	28.1	20.3	19.8	20.6
1TGS	82.1	44.2	70.8	35.2	57.5	36.1	60.6	28.2	32.2	26.9	25.7	27.4	22.5
2TGP	88.7	15.4	44.3	40.0	65.2	11.6	43.2	36.1	26.8	25.2	19.4	21.2	19.8
4TPI	44.5	0.0	4.1	19.6	40.5	0.0	4.1	14.2	23.1	23.8	27.3	26.4	20.9
4SGB	87.1	17.9	22.0	7.3	61.4	12.7	17.3	5.4	26.5	25.6	23.1	23.4	18.7
2KAI	0.8	29.8	3.9	38.7	0.6	20.2	3.4	42.6	25.9	29.1	23.1	20.1	20.1
1CHO	17.2	36.2	29.4	19.1	39.1	35.8	17.2	20.5	26.4	29.6	25.8	19.6	60.5
2SEC	77.1	17.2	23.4	65.1	45.1	11.5	13.5	51.9	33.3	29.2	33.1	24.4	19.5
2SNI	69.8	20.0	77.9	74.1	56.1	19.6	61.2	62.3	26.8	20.2	25.7	24.0	21.5
1TEC	88.5	21.0	83.7	21.9	55.7	20.7	63.8	17.2	33.2	20.7	27.1	26.4	20.9
2CPK	36.8	53.2	83.3	52.6	44.9	63.6	76.5	63.2	38.1	39.1	41.1	38.7	46.6
4CPA	57.4	14.4	93.0	48.2	50.8	8.9	75.6	31.3	24.1	27.5	27.5	33.0	21.3
1FDL	36.4	30.8	5.7	26.6	29.2	24.8	4.7	19.6	14.6	12.1	11.8	13.4	11.7
2HFL	11.3	5.7	19.7	64.9	12.6	5.6	41.2	40.2	20.5	12.9	20.1	18.7	12.7
3HFM	9.7	18.5	0.0	28.5	6.5	23.7	0.0	46.6	13.5	18.1	15.8	13.2	3.1
1NCB	9.1	21.1	0.0	7.4	11.3	19.9	0.0	5.4	2.9	3.6	4.1	8.2	3.1
2HHR	58.8	23.5	37.2	32.3	13.6	8.9	27.8	23.8	3.5	4.6	2.4	5.8	2.5
Average	46.5	24.3	38.8	33.3	35.6	21.7	33.6	29.4	23.4	22.0	21.9	21.4	20.3
SD	31.0	13.8	32.2	21.1	21.2	15.7	27.4	18.8	9.5	9.1	9.3	8.1	14.1
Group mean		35.8				30.1				22.2			
Group SD		26.4				21.4				8.9			

Table 6. *Ranking by geometry*

PDB	Rank	Degeneracy
1TPA	4	17
2PTC	4	12
1TGS	1	2
2TGP	4	17
4TPI	3	22
4TPI:Val-Val	1	3
4SGB	1	2
2KAI	5	10
1HNE	2	1
2EST	1	1
1CHO	1	1
2SEC	2	2
2SNI	3	2
1TEC	2	4
2ER9	1	1
5APR	4	14
6TMN	5	18
1TLP	4	9
1CPK	5	16
4HVP	1	1
4CPA	2	1
LYZ:1FDL	1	2
LYZ:2HFL	2	10
LYZ:3HFM	3	10
Average	2.4	6.8
SD	1.4	6.6

Table 7. *Occurrence of amino acids at the binding interface*[a]

Amino acid	Total, %	Ligand, %	Target, %	Ligand %:target %
Cys	2.8 (1.0)	5.5 (2.0)	4.2 (1.5)	1.3
Met	1.2 (0.7)	0.3 (0.2)	1.0 (0.6)	0.3
Phe	3.4 (1.0)	2.4 (0.7)	4.0 (1.1)	0.6
Ile	4.9 (1.1)	6.1 (1.3)	4.6 (1.0)	1.3
Leu	6.7 (0.9)	4.0 (0.5)	5.5 (0.7)	0.7
Val	7.2 (1.0)	4.3 (0.6)	4.9 (0.7)	0.9
Trp	2.0 (1.8)	4.0 (3.6)	1.7 (1.5)	2.4
Tyr	4.6 (1.3)	8.2 (2.4)	4.3 (1.2)	1.9
Ala	7.0 (0.8)	5.2 (0.6)	4.3 (0.5)	1.2
Gly	9.2 (1.0)	11.6 (1.3)	15.6 (1.7)	0.7
Thr	7.7 (1.3)	7.3 (1.2)	9.7 (1.6)	0.8
Ser	10.6 (1.5)	8.2 (1.2)	11.6 (1.6)	0.7
Asn	5.5 (1.3)	7.6 (1.7)	5.4 (1.2)	1.4
Gln	3.8 (1.0)	1.2 (0.3)	3.0 (0.8)	0.4
Asp	4.9 (0.9)	4.6 (0.8)	7.2 (1.3)	0.6
Glu	3.7 (0.6)	2.1 (0.3)	2.3 (0.4)	0.9
His	1.5 (0.7)	1.2 (0.6)	3.5 (1.7)	0.3
Arg	3.5 (0.8)	6.4 (1.4)	2.4 (0.5)	2.7
Lys	5.0 (0.7)	3.0 (0.4)	2.1 (0.3)	1.5
Pro	4.8 (1.0)	6.7 (1.5)	2.8 (0.6)	2.4

[a] All percentages are for the binding interface except those in column 2, which represent the entire protein. The ratios of observed percentages to their averages found for globular proteins (Table 1-1 of Creighton, 1984) are shown in parentheses.

condition may reflect the crudeness of the C^{α} model proposed in this analysis for precisely identifying binding sites. The results indicate, however, that this method can be used to limit the correspondence between predicted and actual binding sites to a few well-identified regions that may be subjected to further analysis with a more complete atomic model. An alternative possibility that may relate to the imperfect correspondence between the more hydrophobic clusters and the binding interface may be that another ligand, not yet identified, binds to this site. The results for lysozyme support this view. Lysozyme is known to interact with at least 3 different antibodies at 3 different sites. The rankings for lysozyme reflect these interactions and find correspondence with the first, fourth, and fifth most hydrophobic regions with the known antibody-binding sites. In the absence of a known ligand attached to a strongly hydrophobic site, one also cannot rule out the possibility that this site is involved in contact with another molecule within the crystallographic unit cell. An additional consideration suggests that for some cases nonhydrophobic interactions play a strong role in ligand attachment. The case of the anti-lysozyme Fab' molecule (1FDL) is such an example. Hydrogen bonding of 1FDL to Lys 116 and Asp 119 of lysozyme provide the antigenic determinant of this interaction (Fischmann et al., 1991). Measures of hydrophobicity could not detect these interactions and hence the inability of our procedure to detect this binding interface. Taken together, these considerations emphasize the need to examine carefully the atomic details of amino acid clusters identified with the simple procedure presented here, to assess further the role of properties other than hydrophobicity at the binding interface (Tello et al., 1993).

Although this procedure has been validated on test cases involving primarily protein–protein interactions, identification of important hydrophobic contacts involving non-peptidic ligands may also be possible. Application of our method to dihydrofolate reductase (4DFR) found the methotrexate-binding site as the sixth-ranked position of the hydrophobic clusters. This result suggests that hydrophobic interactions play a rather general role in the interaction between a target protein and any ligands. Efforts to design rational compounds on the basis of such interactions have recently been attempted in the case of steroid-based ligands (Kellogg et al., 1991).

Applications

The success of this approach for identifying ligand attachment sites may have applicability to cases where little or no information is available about the position of the bound ligand. The following examples illustrate this point. Two cases where proteins are known to form complexes with peptides or other proteins are the human histocompatibility antigen HLA-A2 (3HLA) (class I encoded in the major histocompatibility complex [MHC]) and the fragment CD4 (2CD4) receptor. Both of these molecules are of considerable therapeutic interest as potential targets for immunological modulation. The analysis of HLA considered the α_1 and α_2 domains (residues 1–182), which are the 2 parallel α-helices atop a β-sheet forming a cleft as the expected antigen-binding site (Saper et al., 1991). The analysis of CD4 considered the N-terminal fragment comprising 2 domains (V1, V2) of the 4 calculated immunoglobulin-like extracellular domains (Ryu et al., 1990; Wang et al., 1990). This portion of CD4 is reported to be a receptor for the HIV gp120 fragment

of the HIV coat protein (Ryu et al., 1990; Wang et al., 1990). The analysis of the surfaces of these 2 molecules follows.

For HLA, 1,773 hydrophobic clusters are found for the lattice model of the antigen-binding domains referred to as $\alpha_1\alpha_2$. The strongest hydrophobic clusters are associated with the cleft formed by $\alpha_1\alpha_2$ and correspond to the position of extra electron density found in the crystal structure (Saper et al., 1991). The cleft itself is convexly curved and narrows at the middle of its lengthwise center. The 2 strongest ranked hydrophobic clusters include the so-called anchor residues for the termini of different length peptides found to bind the cleft with varying degrees of bulge in their middle (Guo et al., 1992; Parham, 1992; Madden et al., 1993). These results are consistent with the binding position for 2 (Fremont et al., 1992) and 5 (Madden et al., 1993) different viral peptides that have been solved crystallographically for murine H-2K^b and human HLA-A2 complexes, respectively. Table 8A lists residues observed in the murine H-2K^b structure that contain atoms observed with van der Waals contacts to the pair of peptides refined by Fremont et al. and residues from our calculation based on the 3 strongest hydrophobic clusters. Strong agreement is found between these lists as indicated by the common residues in 13 of the 24 amino acid positions observed to be in contact with the co-crystallized peptides. The results for HLA-A2 (Madden et al., 1993) are consistent with those for H-2K^b (Fremont et al., 1992), particularly with respect to contacts with the peptide termini. Among the additional contacts reported for HLA-A2 interaction with viral peptides (cf. Table 1, Madden et al., 1993), those involving Tyr 59 and Tyr 123 with the N- and C-termini of the tested peptides, respectively, are also predicted by hydrophobicity analysis to form contacts (cf. Table 8A). These results suggest that a portion of the MHC class I surface interacting with the termini of a candidate peptide may be determined on the basis of hydrophobicity.

Another cluster of slightly lower hydrophobicity is related to the bend of the cleft. This bend forms a concave surface under the cleft. Three strong hydrophobic clusters are associated with this concavity, 2 of which are at the interface of the $\alpha_1\alpha_2$ and β_2m domains. Table 8B lists $\alpha_1\alpha_2$ residues of these 2 and compares them to residues observed at the interface with β_2m (Saper et al., 1991). As in the previous case, there is agreement between residues observed at the β_2m interface and those calculated on the basis of hydrophobicity. An additional comparison can be made between the entries in Table 8A and B by noting residues common to both calculated lists. For example, Phe 9, Val 25, and Tyr 27 appear in both lists and suggest a possible dual role where these residues might participate in β_2m and viral peptide interactions. Their appearance in both lists may be consistent with the ternary model where high-affinity peptide-binding sites are generated by the interaction of β_2m with the class I MHC heavy chain (Boyd et al., 1992). Such an interaction might be mediated by the strongly hydrophobic residues found in these 3 positions.

The total number of hydrophobic clusters for CD4 is 1,728. The best hydrophobic cluster exhibits considerable overlap and can be grouped into 1 larger cluster. This cluster is associated with 64.4% of the total accessible surface of the CD4 2-domain fragment, nearly twice that found for the HLA domains. This large fraction of the surface area calculated to be involved in binding is consistent with studies of interactions between CD4 and gp120, monoclonal antibodies, and class II MHC molecules (Clayton et al., 1989; Ryu et al., 1990; Wang et al., 1990; Capon & Ward, 1991; Szabo et al., 1992). Limited competition for binding to CD4 is found for these proteins, the exception being some competition between the OKT4 antibody set and the MHC molecule. The findings from analysis of the surface hydrophobicity are consistent with these published results and indicate an unusually large fraction of the surface of CD4 to be potentially receptive to other molecules.

The best protein clusters calculated for the CD4 2-domain (V1, V2) fragment identify 3 parts of the receptor as the most favorable binding areas (Fig. 2). The nomenclature of strand labeling found in Capon and Ward (1991) and Wang et al. (1990) is used to report the results. The strongest area (area I, shown in red in Fig. 2) on the receptor is at the interdomain (Wang et al., 1990) connection of the first (V1: residues 1–98) and second (V2: residues 99–176) domains. In Figure 2, the upper half of the molecule is domain 1 (V1) and the lower half, domain 2 (V2). The residues composing area I are 4–10, 12–14, 75–77, 99–102, 105–106, 119, 129, 160–162, 165–170, and 172–174. The strands of these residues are mainly V1: A, A–B, B, and E–F, and V2: A, F, and G. Thus, the corresponding template wraps almost completely around the interdomain connection. A second strong area (area II, shown in green in Fig. 2) is at the end of the rodlike fragment on domain 2. Area II residues are 136, 138–140, 143–150, and 152 (V2: C0, E, and E–F). The third area (area III, shown in purple) is at the other end of the fragment

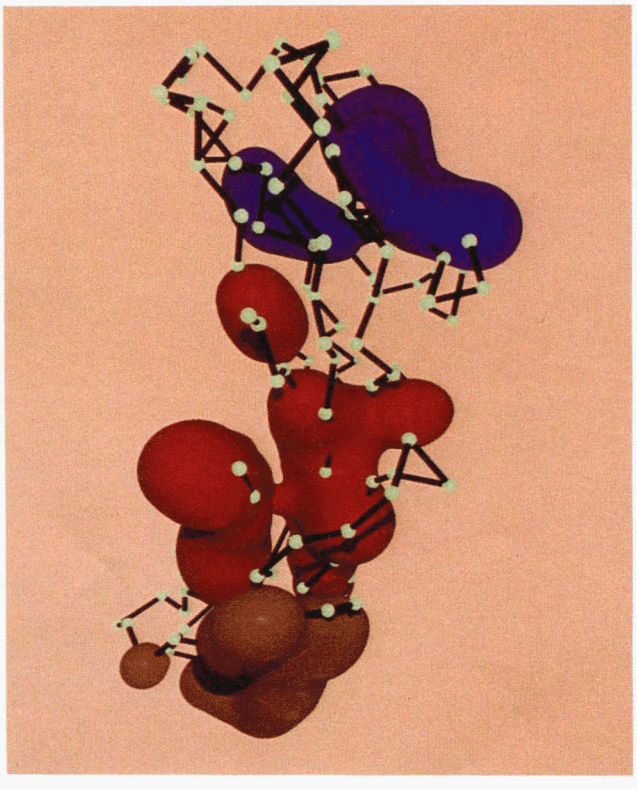

Fig. 2. A C^α tracing of the N-terminal fragment of CD4 (Ryu et al., 1990; Wang et al., 1990). The upper part of the molecules is domain 1 (V1), and the lower part is domain 2 (V2). Residues composing the groups of clusters, labeled area I (red), area II (green), and area III (purple), are represented by spheres centered on the C^α positions with expanded radii of 4.1 Å.

Table 8. *Comparison of observed and calculated data for HLA and CD4*

Observed	Calculated	Observed	Calculated	Observed	Calculated
		A. H-2K(b)[a]			
Tyr 7	**Tyr 7**	**Phe 74**	**His 74**		Ala 140
Val 9	**Phe 9**	**Asp 77**	**Asp 77**		Thr 142
	Val 25		Leu 78	**Thr 143**	**Thr 143**
	Gly 26	**Thr 80**	**Thr 80**	Lys 146	
	Tyr 27	**Leu 81**	**Leu 81**	Trp 147	
	Val 34	Tyr 84		Glu 152	
	Arg 35		Val 95	Arg 155	
	Met 45	Ser 99		Leu 156	
	Ile 52	Gln 114		Tyr 159	
	Tyr 59	**Tyr 116**	**Tyr 116**	**Thr 163**	**Thr 163**
Glu 63	**Glu 63**		Ala 117		Cys 164
Lys 66	**Lys 66**		Tyr 118	**Trp 167**	**Trp 167**
	Val 67		Tyr 123		Leu 168
Asn 70			Ile 124		Arg 170
Ser 73			Ala 139	**Tyr 171**	**Tyr 171**
		B. HLA[b]			
	His 3	**Val 25**	**Val 25**		Asp 102
	Ser 4	**Gly 26**	**Gly 26**		Val 103
	Met 5	**Tyr 27**	**Tyr 27**		Arg 111
	Arg 6	**Thr 28**	**Val 28**		Gly 112
	Tyr 7	**Leu 30**	**Asp 30**		Tyr 113
Phe 8	**Phe 8**	**Gln 32**	**Gln 32**	Gln 115	
Phe 9	**Phe 9**	Arg 35		Tyr 116	
Thr 10	**Thr 10**	Arg 48		Aal 117	
Val 12		**Thr 94**	**Thr 94**	Asp 119	
	Phe 22	**Gln 96**	**Gln 96**	Gly 120	
Ile 23	**Ile 23**		Gly 100	Lys 121	
	Ala 24		Gly 101	Asp 122	
		C. CD4[c]			
Lys 29	**His 27**	**Phe 43**	**Phe 43**	**Arg 59**	**Arg 59**
	Lys 29	**Leu 44**	**Leu 44**		Ser 60
	Ile 34	**Thr 45**	**Thr 45**	Gln 64	
	Lys 35	**Gly 47**	**Gly 47**	Phe 67	
	Ile 36	Ser 49		Glu 77	
Gly 38	**Gly 38**	Asn 52		Thr 81	
Gln 40	**Gln 40**	Ala 55		Glu 85	
Gly 41			Asp 56	Glu 87	
Ser 42			Ser 57	Asp 88	
		Arg 58	**Arg 58**	Gln 89	

[a] Comparison of H-2K(b) $\alpha(1)\alpha(2)$ residues observed within van der Waals contacts to 2 different viral peptides (Fremont et al., 1992) with the calculated 3 most hydrophobic clusters. Boldface type indicates agreement between predicted and observed residues.

[b] Comparison of HLA $\alpha(1)\alpha(2)$ residues containing atoms observed at the interface with $\beta(2)$ microglobulin (Saper et al., 1991) to those of the calculated strongest hydrophobic cluster outside the cleft. Boldface type indicates agreement between predicted and observed residues.

[c] Comparison of CD4 residue mutations (Ryu et al., 1990) affecting gp120 binding to the residues in the calculated protein clusters composing area III in Figure 3. Boldface type indicates agreement between predicted and observed residues.

in domain 1, residues 27, 29, 34–36, 38, **40**, **43–45**, **47**, and 56–60 (V1: C–C′, C′, C′–C″, D, and D–E). This portion of the surface corresponds to the predictions for gp120 binding; in particular, the residues in boldface are in the exposed C′C″ ridge (residues 40–55) considered critical for gp120 interaction (Capon & Ward, 1991). In Table 8C, we list the residues whose mutations affect gp120 binding (Ryu et al., 1990) and those forming the clusters in area III. A strong correspondence is found between the residues common to both lists.

The results of our analysis may also have applicability to the design of therapeutic agents. The results reported here for the HIV-1 protease surface provide an example. The present algo-

rithm's identification of multiple clusters for 4HVP with comparably strong hydrophobicity suggests alternative targets for drug design. Such sites have been referred to as exosites. The analysis of the apo-enzyme form of the dimer reveals 2 possible alternate binding targets. One region is located at the base of the flap, described as the cantilever for flap motion (Harte et al., 1990). The other region extends from the active site to the flap. Attachment of a peptide to either of these regions might inhibit flap motion and possibly influence access to the catalytic triad.

Materials and methods

Selection of proteins

The surface hydrophobicity of 38 crystal structures of proteins comprising 9 different enzyme types, 7 antibodies or antibody fragments, and 3 additional complexes (Table 1A,B) are examined. These proteins have been selected because the structure of the protein complexed with protein, peptide, or nonstandard peptide ligands has been solved and is available for purposes of comparison with the calculation. Even though many of the enzyme molecules are quite similar to others within this group, there is a substantial range of diversity. All but 1 of these enzymes are cataloged as hydrolases according to their EC class name (Lehninger, 1975), the exception being 2CPK, which belongs to the transferase class. These enzymes can be further grouped according to families specified by amino acid motifs at their sites of interactions with substrate, co-factor, or hapten (Bairoch, 1992). Five enzymes (1TPA, 2PTC, 1TGS, 2TGP, and 4TPI) are similar in length and contain the active site histidine and serine residues characteristic of the trypsin family of serine proteases. These 5 enzymes share identical sequences but differ in their coordinate positions. The C^{α} positions of 2PTC, 1TGS, 2TGP, and 4TPI, when compared to 1TGS, are within 0.1 Å, 1.2 Å, 0.9 Å, and 1.1 Å RMS deviation from each other, respectively. Serine proteinase B (4SGB) also contains both active site residues but is shorter in length than the other members of this family and shares only 20% sequence identity with 1TGS. One enzyme (2KAI) contains only the active site serine of this family and demonstrates 41% and 20% sequence identity to the 5 longer members of the trypsinogen family and 4SGB, respectively. Two elastases appear as members of the serine protease family with 2EST containing both active site residues, whereas 1HNE contains only the active site histidine. This pair of enzymes share 44% sequence identity. 1CHO is included as a chymotrypsin member of the serine protease family because it contains the active site serine. Three enzymes (2SEC, 2SNI, and 1TEC) fall in the subtilase family of serine proteases as identified by the Asp, His, Ser triad. The sequence identity among this group ranges from 70 to 44%. 2ER9 and 5APR are members of the eukaryotic aspartyl protease family, share 40% sequence identity, and have RMS deviations between their C^{α} positions of greater than 9 Å. The zinc-binding region signature identifies 6TMN and 1TLP as zinc-metallopeptidases and 4CPA as a zinc-carboxypeptidase. One member of the eukaryotic protein kinase family (2CPK) and 1 acid proteinase (4HVP) complete the set of enzymes examined. Six complexes fall into the class of antibody–antigen interactions. Three of these are Fab′ fragments directed against the lysozyme antigen (1FDL, 2HFL, and 3HFM). These 3 fragments share 75–77% sequence identity. The other 3 Fab′ fragments studied are directed against a synthetic octamer (1HIM), a fragment of myohemerythrin (2IGF), and neuraminidase (1NCB). One complex involves an interaction between the complement-binding antibody fragment (FC) and fragment B of protein A. The 3 remaining complexes studied include hirudin:α-thrombin (1HTC), fragments of growth hormone:growth hormone receptor (2HHR), and the retinol binding protein: retinol complex (1RBP).

Geometry

Any strategy to search a molecular surface requires tradeoffs between the accuracy of the surface representation and the requirements to search the surface rapidly and completely. Simplified models such as the C^{α} backbone (Levitt, 1976) selected for this analysis have been used widely to examine the details of protein stability, packing, and folding (Levitt, 1976; Covell & Jernigan, 1990; Dill, 1990; Skolnick & Kolinski, 1990). A further simplification can be made in searching the space around the protein by mapping the C^{α} positions to a regular lattice. Covell and Jernigan (1990) found that, among several lattice types tested, the one providing the best fits of virtual bond and torsion angle geometry in the C^{α} backbone was the face-centered cubic lattice of edge dimension 3.8 Å, corresponding to the fixed virtual bond distance between neighboring C^{α}'s. With this lattice, an accurate model of the C^{α} positions of the target molecule is obtained, with overall fits of about 1 Å RMS deviation from the crystallographic protein structure.

Another important feature of lattice models is their ability to easily provide positions exterior to the target molecule upon which a substrate might be placed. This is accomplished by simply extending the same lattice used to define the C^{α} backbone into the region surrounding the protein (Jernigan et al., 1989). This grid then defines a *shell* of positions exterior to the molecular surface that can be examined for their interactions with nearby residues on the target molecule. The thickness of the shell is established by removing lattice points that are either too close to or too far away from the protein. The outer boundary of the shell includes points closer than 9.0 Å from any protein residue, whereas the inner boundary includes only lattice points farther than 6.1 Å from any protein residue (Fig. 3). The inner boundary is based on observation of the crystal complexes studied; the outer boundary is the limit for which the contact energy parameters are likely to be valid (Christenson & Claesson, 1988). These exterior positions can be thought of as defining the portion of a molecule's surface that would be accessible to an approaching protein (represented by its C^{α} coordinates) and thus parallels the concept of molecular surfaces based on their accessibility to a water molecule probe. The set of protein residues within a sphere of radius 9.0 Å from each shell point defines the nearest protein neighbors of each shell point. This set of residues is referred to as a *cluster*. Clusters are then ranked according to their total hydrophobicity calculated as explained below. The geometry of this system is described by the example in Figure 3. In this figure, the connected dots represent a portion of the protein under investigation. The hatched dot is an exterior shell point, which defines a cluster of protein residues to be those amino acids located within a fixed distance of this exterior point. Clusters, represented by the connected open dots in this figure, are formed for the entire protein and ranked in terms of their total hydrophobicity using an acceptable scale. The process

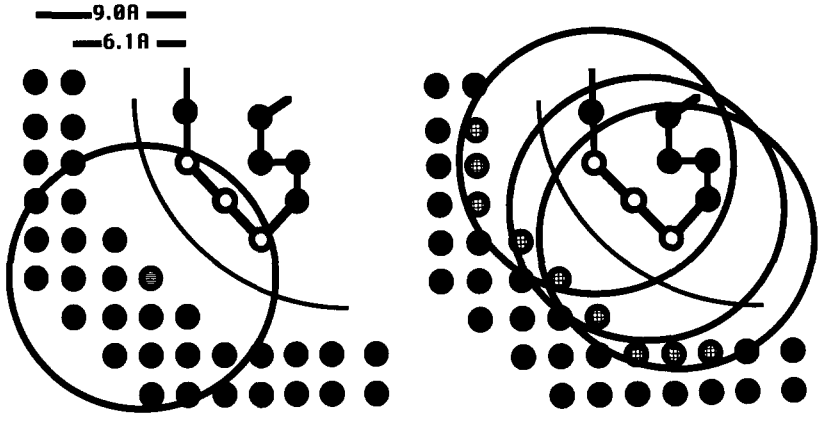

Fig. 3. Two-dimensional illustration of the procedure for defining protein clusters. A shell of lattice points surrounding the protein is chosen to include positions less than 9.0 Å and greater than 6.1 Å from the protein. The outer dimension is the appropriate limit for which the contact energy parameters (Christenson & Claesson, 1988) are valid. The inner dimension is based on near-neighbor statistics for the crystal complexes studied. An arc separates the protein residues (connected dots) from the surrounding lattice points. The circle centered on the hatched lattice point defines the near-neighbor protein residues composing a cluster. In this example, 3 residues are included in the cluster. The hydrophobicity of this cluster is the sum of the hydrophobicities for its constituent amino acids.

of fitting the coordinates to the lattice, generating the shell of surrounding lattice points, and identifying all clusters can be completed in less that 2 CPU min on a Silicon Graphics 310 Workstation.

Determination of cluster hydrophobicity

The hydrophobicity of each protein cluster $\Phi_{cluster}$ is taken to be the sum of the hydrophobicities e_i of its residues: $\Phi_{cluster} = \sum_{i=1}^{N} e_i$, where N is the number of residues in the cluster. These values are based on the contact energies calculated by Miyazawa and Jernigan (1985). Their statistical study of nearest-neighbor residue contacts in protein structures in the Brookhaven Protein Data Bank (PDB) (Bernstein et al., 1977; Abola et al., 1987) determined local neighborhoods of the 20 residue types in 42 high-resolution X-ray crystal structures and used a quasichemical model of the pairwise interactions between types of amino acids and solvent to calculate a set of residue-residue contact energies e_{ij}. These pairwise contact energies were derived for a model using a single point representation of the amino acids in which the point is placed at the center of its side chain atom positions and all interactions within a distance of 6.5 Å are counted (Miyazawa & Jernigan, 1985).

Averages of these pairwise contact energies are used as the hydrophobicities for each residue type, i:

$$e_i = \frac{\sum_{j=1}^{20} e_{ij} n_{ij}}{\sum_{j=1}^{20} n_{ij}}, \qquad (3)$$

where the contact energy e_{ij} is the Miyazawa and Jernigan (1985) derived energy difference between an ij amino acid pair and the respective amino acid-solvent pairs. n_{ij} is the number of ij contacts in the set of structures used to calculate e_{ij}. These contact energies can be understood in terms of the hydrophobic-hydrophilic designations of amino acids and the pairings that contribute most to protein stability (Miyazawa & Jernigan, 1985; Covell & Jernigan, 1990; Covell, 1992, 1994). The ranking of these terms is:

strongest	hydrophobic-hydrophobic
intermediate	hydrophobic-hydrophilic
weakest	hydrophilic-hydrophilic

Thus, each of the 20 naturally occurring amino acids is assigned a hydrophobicity.[7] This scale shows a strong correlation with the Tanford-Nozaki scale (Nozaki & Tanford, 1971) and others, as shown by Cornette et al. (1987). Determination of surface hydrophobicity for all clusters on proteins comparable in size to those studied here can be completed in less than 1 CPU min on a Silicon Graphics 310 Workstation.

Calculation of buried surfaces

A quantitative definition of the molecular interface in a crystal complex is the surface area buried by the bound ligand. Comparisons can be made between this surface and the hydrophobic clusters identified in our analysis by simply comparing the portion of the surface buried by the ligand with that buried by the set of exterior grid points associated with a cluster. This set of exterior grid points represents those points that, if occupied by an approaching ligand, have access to the amino acid residues in a cluster (Fig. 3).

The accessible surface area calculations are performed on the proteins using the Connolly molecular surface program (Connolly, 1983a, 1983b) with a 1.4-Å-radius probe. Surface residues are defined as those with areas greater than or equal to 10.0 Å^2. This threshold is based on the finding (Miller et al., 1987) that the maximum variation in surface areas of most amino acid residues among several structures of the same macromolecule is less than 10.0 Å^2. An average residue radius of 2.6 Å was used for calculating the surface buried by each set of exterior grid points. This value represents an average of residue centroid-centroid distances in the 42-protein set used by Miyawaza and Jernigan (1985).

Acknowledgments

We thank Prof. Ruth Nussinov and Dr. Stan Burt for useful discussions and suggestions regarding preparation of the manuscript, Mr. Gary W. Smythers for valuable assistance in the assignment of enzymes to their respective families, and Dr. Kai-Li Ting for her assistance in analysis of centroid-centroid distances. We thank the Intramural Targeted Antiviral AIDS program for support for this project. Special thanks go to the staff of the Biomedical Supercomputing Center, FCRDC, Frederick, Maryland, for their assistance and for access to the Cray Y-MP su-

[7] The values of e_i are: F −5.12, M −4.91, I −4.88, L −4.65, W −4.36, V −4.17, C −4.00, Y −3.24, A −2.82, H −2.75, G −2.34, T −2.30, P −2.22, R −2.18, S −2.07, Q −1.98, E −1.94, N −1.90, D −1.81, K −1.50.

percomputer. This research is sponsored in part by the National Cancer Institute, DHHS, under contract N01-CO-74102 with Program Resources, Inc. The contents of this publication do not necessarily reflect the views or policies of the DHHS, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government.

References

Abola EE, Bernstein FC, Bryant SH, Koetzle TF, Weng J. 1987. Protein Data Bank. In: Allen FH, Bergerhoff G, Sievers R, eds. *Crystallographic databases—Information content, software systems, scientific applications*. Bonn/Cambridge/Chester: Data Commission of the International Union of Crystallography. pp 107–132.

Bairoch A. 1992. PROSITE: A dictionary of sites and patterns in proteins. *Nucleic Acids Res 20*:2013–2018.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J Mol Biol 112*:535–542.

Bethe AD. 1993. Density-functional thermochemistry. III. The role of exact exchange. *J Chem Phys 98*:5648–5652.

Boyd LF, Kozlowski S, Margulies DH. 1992. Solution binding of an antigenic peptide to a major histocompatibility complex class I molecule and the role of β-microglobulin. *Proc Natl Acad Sci USA 89*:2242–2246.

Brooks CL III, Karplus M, Pettitt B. 1988. *Proteins: A theoretical perspective of dynamics, structure, and thermodynamics*. New York: J. Wiley.

Capon DJ, Ward RHR. 1991. The CD4-gp120 interaction and AIDS pathogenesis. *Annu Rev Immunol 9*:649–678.

Cherfils J, Duquerroy S, Janin J. 1991. Protein–protein recognition analyzed by docking simulation. *Proteins Struct Funct Genet 11*:271–280.

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

Christenson HK, Claesson PM. 1988. Cavitation and the interaction between macroscopic surfaces. *Science 239*:390–392.

Clayton LK, Sieh M, Pious DA, Reinherz EL. 1989. Identification of human CD4 residues affecting class II MHC versus HIV-1 gp120 binding. *Nature 339*:548–551.

Connolly ML. 1983a. Analytical molecular surface calculation. *J Appl Crystallogr 16*:548–558.

Connolly ML. 1983b. Solvent accessible surfaces of proteins and nucleic acids. *Science 221*:709–713.

Cornette JL, Cease KB, Margalit H, Spouge JL, Berzofsky JA, DeLisi C. 1987. Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J Mol Biol 195*:659–685.

Covell DG. 1992. Folding protein α-carbon chains into compact forms by Monte Carlo methods. *Proteins Struct Funct Genet 14*:192–204.

Covell DG. 1994. Lattice model simulations of polypeptide chain folding. *J Mol Biol 235*:1032–1043.

Covell DG, Jernigan RL. 1990. Conformations of folded proteins in restricted spaces. *Biochemistry 29*:3287–3294.

Creighton TE. 1984. *Proteins*. New York: W.H. Freeman and Company.

Danziger DJ, Dean PM. 1989a. Automated site-directed drug design: A general algorithm for knowledge acquisition about hydrogen-bonding regions at protein surfaces. *Proc R Soc Lond B 236*:101–113.

Danziger DJ, Dean PM. 1989b. Automated site-directed drug design: The prediction and observation of ligand point positions at hydrogen-bonding regions on protein surfaces. *Proc R Soc Lond B 236*:115–124.

Dill KA. 1990. Dominant forces in protein folding. *Biochemistry 29*:7133–7155.

Fischmann TO, Bentley GA, Bhat TN, Boulot G, Mariuzza RA, Phillips SEV, Tello D, Poljak RJ. 1991. Crystallographic refinement of the three-dimensional structure of the Fab d1:3-lysozyme complex at 2.5 Å resolution. *J Biol Chem 266*:12915–12920.

Fremont DH, Matsumura M, Stura EA, Peterson PA, Wilson IA. 1992. Crystal structures of two viral peptides in complex with murine MHC class I H-2K^b. *Science 257*:919–927.

Goodford, P. 1985. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J Med Chem 28*:849–857.

Guo H, Jardetzky TS, Garrett TPJ, Lane WS, Strominger JL, Wiley DC. 1992. Different length peptides bind to HLA-Aw68 similarly at their ends but bulge out in the middle. *Nature 360*:364–366.

Harte WE Jr, Swaminathan S, Mansuri MM, Martin JC, Rosenberg IE, Beveridge DL. 1990. Domain communication in the dynamical structure of human immunodeficiency virus 1 protease. *Proc Natl Acad Sci USA 87*:8864–8868.

Horton N, Lewis M. 1992. Calculation of the free energy of association for protein complexes. *Protein Sci 1*:169–181.

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

Jernigan RL, Margalit H, Covell DG. 1989. Coarse graining conformations: A peptide binding example. In: Beveridge DL, Lavery R, eds. *Theoretical biochemistry and molecular biophysics*. Schenectady, New York: Adenine Press. pp 69–76.

Johnson BG, Gill PMW, Pople JA. 1993. The performance of a family of density functional methods. *J Chem Phys 98*:5612–5626.

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

Kelley RF, O'Connell MP. 1993. Thermodynamic analysis of an antibody functional epitope. *Biochemistry 32*:6828–6835.

Kellogg GE, Semsus SF, Abraham DJ. 1991. Hint: A new method of empirical hydrophobic field calculation of comfa. *J Comput Aided Mol Design 5*:545–552.

Knighton DR, Zheng JZ, Ten Eyck LF, Xuong N, Taylor SS, Sowadski JM. 1991. Structure of a peptide inhibitor bound to the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science 253*: 414–420.

Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE. 1982. A geometric approach to macromolecular-ligand interactions. *J Mol Biol 161*:269–288.

Lehninger AL. 1975. *Biochemistry, 2nd ed*. New York: Worth.

Levitt M. 1976. A simplified representation of protein conformations for rapid simulation of protein folding. *J Mol Biol 104*:59–107.

Madden DR, Garboczi DN, Wiley DC. 1993. The antigenic identity of peptide-MHC complexes: A comparison of the conformations of five viral peptides presented by HLA-A2. *Cell 75*:693–708.

Miller S, Janin J, Lesk AM, Chothia C. 1987. Interior and surface of monomeric proteins. *J Mol Biol 196*:641–656.

Miyazawa S, Jernigan RL. 1985. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules 18*:534–552.

Nicholls A, Sharp KA, Honig B. 1991. Protein folding and association: Insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins Struct Funct Genet 11*:281–296.

Nozaki Y, Tanford C. 1971. The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. *J Biol Chem 246*: 2211–2217.

Parham P. 1992. Deconstructing the MHC. *Nature 360*:300–301.

Privalov PL, Gill SJ. 1988. Stability of protein structure and hydrophobic interaction. *Adv Protein Chem 39*:191–234.

Rullmann JAC, van Duijen PT. 1990. Potential energy models of biological macromolecules. A case for ab initio quantum chemistry. *Rep Mol Theory 1*:1–21.

Ryu S, Kwong PD, Truneh A, Porter TG, Arthos J, Rosenberg M, Dai X, Xuong N, Axel R, Sweet RW, Hendrickson WA. 1990. Crystal structure of an HIV-binding recombinant fragment of human CD4. *Science 348*:419–426.

Saper MA, Bjorkman PJ, Wiley DC. 1991. Refined structure of the human histocompatibility antigen HLA-A2 at 2.6 Å resolution. *J Mol Biol 219*: 277–319.

Sharp KA, Nicholls A, Fine R, Honig B. 1991a. Reconciling the magnitude of the microscopic and macroscopic hydrophobic effects. *Science 252*: 106–109.

Sharp KA, Nicholls A, Friedman R, Honig B. 1991b. Extracting hydrophobic free energies from experimental data: Relationship to protein folding and theoretical models. *Biochemistry 30*:9686–9697.

Shoichet BK, Bodian DL, Kuntz ID. 1992. Molecular docking using shape descriptors. *J Comput Chem 13*(2):1–18.

Skolnick J, Kolinski A. 1990. Simulations of the folding of globular proteins. *Science 250*:1121–1125.

Szabo G Jr, Pine PS, Weaver JL, Rao PE, Aszalos A. 1992. CD4 changes conformation upon ligand binding. *J Immunol 149*:3596–3604.

Tanford C. 1980. *The hydrophobic effect: Formation of micelles and biological membranes, 2nd ed*. New York: Wiley and Sons.

Tello D, Goldbaum FA, Mariuzza RA, Ysern X, Schwarz F, Poljak RJ. 1993. Immunoglobulin superfamily interactions. *Biochem Soc Trans 21*:943–946.

Varadarajan R, Connelly P, Sturtevant JM, Richards FM. 1992. Heat capacity changes for protein–peptide interactions in the ribonuclease s system. *Biochemistry 31*:1421–1426.

Wang J, Yan Y, Garrett TPJ, Liu J, Rodgers DW, Garlick RL, Tarr GE, Husain Y, Reinherz EL, Harrison SC. 1990. Atomic structure of a fragment of human CD4 containing two immunoglobulin-like domains. *Science 348*:411–418.