doi:10.1016/j.jmb.2004.06.072 J. Mol. Biol. (2004) 342, 665–679

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Protein–Protein Recognition: Juxtaposition of Domain and Interface Cores in Immunoglobulins and Other Sandwich-like Proteins

Vladimir Potapov[1], Vladimir Sobolev[1]*, Marvin Edelman[1] Alexander Kister[2,3] and Israel Gelfand[2]

[1]*Department of Plant Sciences Weizmann Institute of Science Rehovot 76100, Israel*

[2]*Department of Mathematics Rutgers University, New Brunswick, NJ 08903, USA*

[3]*Department of Health Informatics, University of Medicine and Dentistry of New Jersey, Newark, NJ 07101 USA*

**Corresponding author*

Structural analysis of a non-redundant data set of 47 immunoglobulin (Ig) proteins was carried out using a combination of criteria: atom–atom contact compatibility, position occupancy rate, conservation of residue type and positional conservation in 3D space. Our analysis shows that roughly half of the interface positions between the light and heavy chains are specific to individual structures while the other half are conserved across the database. The tendency for conservation of a primary subset of positions holds true for the intra-domain faces as well. These subsets, with an average of 12 conserved positions and a contact surface of 630 Å^2, delineate the inter- and intra-domain core, a refined instrument with a reduced target for analysis of sheet–sheet interactions in sandwich-like proteins. Employing this instrument, we find that a majority of Ig interface core positions are adjoined in sequence to domain core positions. This was derived independent of geometric considerations, however β-sheet side-chain geometry clearly dictates it. The geometric wedding of the domain and interface cores supports the concept of a rigid-like substructure on the protein surface involved in complex formation and indicates a close relationship between surface determinants and those involved in protein folding of Ig domains. The definitions developed for the Ig interface and domain cores proved satisfactory to extract first-approximation cores for a group of 24 non-Ig sandwich-like proteins, treated as individual structures due to their diverse strand topologies. We show that the same rule of positional connectivity between the rigid domain core and interface core extends generally to sandwich-like proteins interacting in a sheet–sheet fashion. The non-Ig structures were used as templates to analyze sandwich-like interfaces of unresolved homologous proteins using a database merging structure and sequence conservation.

Keywords: protein–protein binding; VL–VH interface; CL–CH1 interface; interlock strands; interlock positions

Introduction

Structural flexibility is crucial for protein folding and function.[1–3] Yet, as Levinthal[4] originally pointed out, an average size protein in transition from random disorder to a uniquely defined structure would need to sort an astronomical number of states of conformational space very quickly. Resolution of this quandary by "hierarchical folding"[5,6] or "nuclear formation"[7,8] is currently being discussed. However, in both approaches, it is assumed that rapid formation of rigid substructures occurs to limit the search space. For example, the structural constraint of two interlocked pairs of β-strands present in all superfamilies of sandwich-like proteins[9] could be a candidate for such substructure. These interlocked pairs stabilize the fold structure[10] and constitute the domain core. Similarly, the presence of relatively rigid elements are implied at protein contacting surfaces, since in a majority of cases only a limited number of rotamer

Abbreviations used: Ig, immunoglobulin; PDB, Protein Data Bank.

E-mail address of the corresponding author: vladimir.sobolev@weizmann.ac.il

changes occur upon protein–ligand binding[11] and only small conformational changes are seen upon protein–protein complex formation.[12]

The rigidity of protein contacting surfaces can be estimated directly by comparing an individual pair of complexed and uncomplexed structures, or indirectly by structural analysis of groups of complexes that share common interfacial features. Many approaches are aimed at directly predicting which residues compose the interface. Schemes were developed to predict a functional residue,[13–16] or set of residues,[17–20] on a protein surface that take part in protein–protein recognition. These approaches are founded on empirical rules derived from an analysis of protein–protein or domain–domain interface features.[21–25] However, it was pointed out that analysis of the binding site of a pair of resolved structures does not necessarily identify the functional protein interface,[26] since not all interface residues form energetically important contacts.[27–29] Thus, the direct approach could miss residues crucial for complex formation. Moreover, prediction of a protein–protein interface is still restricted in accuracy. As a result, intensive efforts are being applied to extract additional levels of information from the database relating to surface–surface interaction. These include: side-chain clustering,[30] secondary structure elements at interfaces,[31] hot spot residues,[28,32–34] residue conservation,[35,36] side-chain conformational entropy,[37] contacting residue pairs,[29,38] empirical potentials,[39] kinetic data[40] and interface analysis at the atomic level.[12,41,42]

The general aim is to clearly distinguish the site of interaction. However, in most cases, predicting a unique interface region has remained elusive, probably because the interface as a whole has approximately the same character as the surface as a whole.[12,24] This realization has led us to seek a statistical approach that could be applied to families of proteins having a significant number of structurally similar, highly resolved members. We reasoned that analysis of a sufficiently large database should permit extraction of a set of universal residue positions (interface "core" positions) crucial for complex formation in such proteins. If properties of such cores differ from those of the interfaces as a whole schemes for the recognition of these surface patches might be more efficient.

In particular, a statistical approach to search for interface core positions might fit the large number of resolved immunoglobulin (Ig) structures in the PDB.[43,44] A start in this direction was made with a small number of Ig proteins, by analyzing β-sheet intra-[45] and inter-domain[46,47] faces. More recently, the use of more extensive databases revealed the regions whose conformations are conserved in almost all variable[48] and constant[49] Ig domains. These geometric conformations underpin the fold structure within the domains.

Here, we use a statistical approach to derive the protein–protein interface core for Ig proteins and analyze its structural conservation. To this end, we evaluate a non-redundant set of 47 Ig structures to determine the residue positions that play a primary role in protein–protein dimer association of the heavy and light chains. We discover the existence of highly conserved positional determinants at the two protein surfaces that presumably allow fast recognition and initial binding of the chains. We further demonstrate that rigidity at the interface surface is geometrically wedded to the domain core, indicating a close relationship between these surface determinants and those involved in protein folding of Ig domains. This close linkage provides an initial basis for extending interfacial structural analysis to 24 non-Ig sandwich-like proteins.

Results and Discussion

Interfaces and interface core

The interface between light chain (L) and heavy chain (H) for a given Ig molecule is composed of the contacting L and H residues, as defined by CSU software,[50] yielding the VL–VH and CL–CH1 interface regions. The procedure for defining the interface core is the following: The virtual interface is first defined as the cumulative set of positions at the interfaces of all 47 complexes in the database. The contact area for a given pair of positions at the virtual interface is calculated as the average contact area between residues at these positions. As a first approximation, the interface core is derived from the virtual interface contact map by taking the minimal number of pair positions forming 80% of the average interface area ("80% core"). The "80% core" is then refined by extracting the primary positions based on several criteria. The first criterion is physical–chemical compatibility of the residue pair, determined by atom–atom contacts.[51] A contact should, in most cases, be attractive (hydrophobic–hydrophobic or hydrophilic–hydrophilic) and formed by side-chain atoms (at least from one partner) to be considered in the interface core. The second criterion is the frequency of a contact position appearing in the data set (high-frequency positions are defined as >40 out of 47). The third criterion is conservation of residue type (hydrophobic, hydrophilic or neutral[48]). The last criterion is positional conservation in 3D space (RMSD of C^{α} atoms ≤ 2.0 Å after superimposition of all 47 structures).

VL–VH interface

The VL–VH interface is formed predominantly of Sheet II residues from VL and VH (Figure 1). Fifty one VL and 46 VH positions compose the virtual VL–VH interface for all 47 Ig structures (Table 1). A clear correlation exists among high-frequency positions, conservation of residue type, and small cluster radius. Superimposition of all 47 structures revealed a tight clustering for about half of the C^{α}

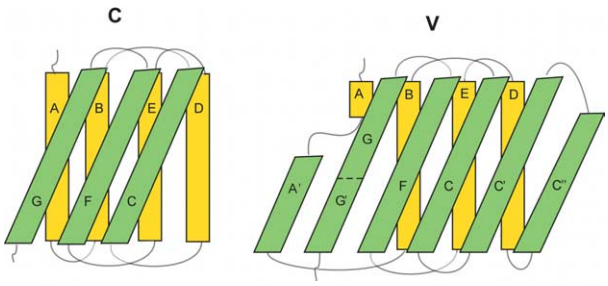

Figure 1. Schematic presentation of the immunoglobulin domains. The constant domain (C) consists of Sheet I (containing strands A, B, E, D) and Sheet II (containing strands C, F, G). The variable domain (V) consists of Sheet I (containing strands A, B, E, D) and Sheet II (containing strands A′, C, C′, C″, F, G, G′). Sheet nomenclature corresponds to Gelfand & Kister[52] with partition of the G strand into G and G′[49] to account for the rotation in the middle of the strand produced by a β-bulge.[48]

atoms (cluster radii <2 Å). For a given structure, 18 to 30 residues (average, 24) from VL and 19 to 27 residues (average, 23) from VH make up the interface, yielding an average interface contact surface of 1074 Å^2 (maximum, 1389 Å^2; minimum, 800 Å^2). The minimal set of contacts, and their residue positions, forming the "80% core" are shown in Figure 2.

The final step in defining the interface core is the determination of the primary positions. These positions are arrived at by elimination: positions C3L, C′10L, F9L and FG1L were eliminated because they are occupied both by hydrophobic and hydrophilic residues, and contact with the FG loop of the VH domain is sporadic; position C″5L was eliminated as it is insufficiently represented at the interface; residues at positions CC′5L, CC′4H and CC′5H were eliminated as they mostly form backbone contacts with their interface partners. The VL–VH interface core is composed of the remaining 11 VL primary positions (C5, C7, C9, CC′6, C′6, C′9, FG15, FG16, FG17 and G6) and the 7 VH ones (C7, C9, CC′6, C′6, F7, G4, G6). The cells corresponding to the contacts between the VL and VH primary positions are shadowed in Figure 2. Hydrophobic contacts are formed primarily between positions C7, CC′6, C′6, F7, G6 in the heavy chain and positions C7, CC′6, F7, FG7, FG8, G6 in the light chain. In most cases, positions C′6, C′9, FG8 and G6 in the light chain also form hydrophobic contacts with the FG loop of VH. Rarely, hydrophilic residues are at these positions (Table 1) but in these cases a strong hydrogen bond is formed. A conserved hydrophilic region of the interface is formed by the interaction of residue side-chains at positions C5 and C9 of the VL domain and the backbone of the FG loop in the VH domain.

Antigen-induced domain rearrangements may occur in an antibody. We analyzed the stability of the interface core and the changes that occur upon antigen–antibody complex formation of Fab 50.1 that exhibits one of the largest conformational changes observed in a single antibody.[53] We compared the VL–VH contact maps of PDB entries 1ggb (uncomplexed) and 1ggi (complexed). While there are approximately 30 changes overall between the two, there are very few changes in interface core positions (C5 and C′9 from the VL domain, which are not in contact with the FG loop of VH in 1ggi, and F7 from VH and C9 from VL, that are not in contact in 1ggb, although other contacts formed by F7 and C9 still exist).

Figure 3 illustrates the relation among the virtual

VH domain

VL domain Pos.	C7	C9	CC'4	CC'5	CC'6	C'6	C"3	C"5	F7	FG	G4	G6	G7
C3										23			
C5										32			
C7										37		22	
C9		39							10				
CC'5									12			10	22
CC'6					21				11			61	
C'6										34	26		
C'9										47			
C'10										23			
C"5										12	13		
F7		10	11	13	33								
F9										17			
FG10										35			
FG15						14	27						
FG16						19		19					
FG17						54				29			
G6	18				53	24				15		13	
G8				11									

Figure 2. VL–VH "80% contact map". The map is derived from the minimal set of residue-residue contacts contributing 80% of the contact surface area to the VL–VH interface. Numbers in the cells are average contact surface areas (total contact surface area in Å^2 for all files at a given position divided by 47). All contacts are formed by Sheet II residues. The cells corresponding to primary hydrophobic–hydrophobic positions are shaded in light gray and those corresponding to hydrophilic–hydrophilic ones, in dark gray. The contacts in the non-shadowed cells are not considered crucial since they are either non-attractive or infrequent (Table 1).

Table 1. Characteristics of the positions contributing to the interface between VL and VH domains

VL domain					VH domain				
Position[a]	Freq.[b]	Contact surface (Å^2)[c]	Cluster radius (Å)[d]	Amino acid composition	Position[a]	Freq.[b]	Contact surface (Å^2)[c]	Cluster radius (Å)[d]	Amino acid composition
oA1	32	10	13.4	$D^{26},E^{12},N^{1},A^{1},Q^{1}$	**A5**	1	0	2.8	$Q^{34},K^{10},H^{1},L^{1},T^{1}$
oA2	2	0	8.8	$I^{26},L^{7},V^{6},A^{2},S^{2},T^{1},N^{1},E^{1}$	**A6**	1	0	1.1	L^{45},V^{1},P^{1}
A5	2	0	2.8	$V^{27},Q^{10},L^{5},A^{2},E^{2},K^{1}$	**C3**	6	2	1.7	$Y^{13},G^{10},W^{10},A^{5},T^{2},S^{2},V^{1},P^{1},N^{1},F^{1}$
A6	1	0	1.5	M^{28},L^{16},V^{3}	**C5**	32	10	0.6	$H^{19},N^{10},S^{9},T^{2},D^{2},E^{1},F^{1},G^{1},Q^{1},Y^{1}$
AA′1	3	0	2.4	$S^{16},A^{13},L^{11},P^{4},F^{1},G^{1},K^{1}$	**C7**	47	18	0.7	$V^{38},I^{5},N^{1},F^{1},A^{1},L^{1}$
BC5	1	0	5.9	$V^{12},S^{11},N^{4},L^{4},G^{3},I^{3},D^{3},F^{2},Y^{2},H^{2},K^{1}$	**C9**	47	52	1.0	Q^{44},K^{2},L^{1}
BC6	7	4	6.9	$N^{11},H^{10},S^{8},T^{5},G^{4},Y^{3},K^{2},F^{1},Q^{1}$	**CC′3**	9	1	4.8	G^{34},E^{12},D^{1}
BC7	3	0	7.7	$S^{18},A^{2},T^{2},N^{1},Y^{1},D^{1}$	**CC′4**	44	23	3.8	$K^{26},Q^{12},N^{3},H^{3},R^{2},G^{1}$
BC8	10	3	7.2	$N^{11},G^{6},S^{3},R^{2},D^{1}$	**CC′5**	47	30	3.5	$G^{31},R^{7},K^{3},S^{3},A^{2},L^{1}$
BC9	4	4	7.3	$G^{12},N^{4},K^{2},H^{1},A^{1},F^{1},T^{1}$	**CC′6**	46	111	1.5	L^{45},F^{1}
BC10	10	3	4.3	$N^{8},Q^{3},S^{2},K^{2},I^{1},D^{1}$	**C′5**	32	1	1.2	E^{43},K^{4}
BC11	1	0	3.3	T^{11},K^{2},I^{1},M^{1}	**C′6**	47	114	0.8	W^{44},F^{1},Y^{1},L^{1}
C3	30	25	1.9	$Y^{30},N^{3},F^{3},T^{2},D^{2},H^{2},A^{2},S^{1},R^{1},K^{1}$	**C′7**	1	0	0.8	$I^{22},V^{15},M^{7},L^{3}$
C4	3	0	0.8	$L^{34},V^{4},A^{4},M^{4},I^{1}$	**C′8**	2	1	1.4	G^{32},A^{14},V^{1}
C5	42	33	0.5	$N^{13},H^{12},A^{9},E^{3},S^{3},T^{2},Y^{2},D^{1},R^{1},K^{1}$	**C′9**	42	16	1.0	$Y^{10},W^{7},R^{6},E^{5},S^{3},T^{3},F^{3},A^{2},V^{2},L^{2},I^{1},Q^{1},D^{1},M^{1}$
C6	1	0	0.5	W^{47}	**C′11**	9	2	1.8	$D^{7},Y^{7},S^{7},N^{6},R^{6},W^{4},I^{2},L^{2},V^{1},F^{1},H^{1}$
C7	47	64	0.6	$Y^{38},V^{3},F^{3},L^{2},H^{1}$	**C″1**	3	1	3.1	$Y^{7},N^{5},G^{5},T^{5},S^{5},E^{4},R^{4},A^{3},D^{3},V^{2},F^{1},K^{1}$
C9	47	51	0.7	Q^{43},E^{3},H^{1}	**C″3**	44	34	1.4	$Y^{11},N^{8},T^{6},E^{5},H^{5},K^{4},F^{2},I^{2},V^{1},G^{1},D^{1},L^{1}$
CC′2	3	1	4.0	P^{37},S^{6},Q^{3},T^{1}	**C″4**	36	7	1.3	$Y^{41},F^{3},N^{1},S^{1},L^{1}$
CC′3	25	4	4.9	G^{35},D^{8},H^{3},N^{1}	**C″5**	43	23	1.2	$N^{18},A^{11},D^{4},S^{4},P^{3},H^{2},L^{2},T^{1},V^{1},G^{1}$
CC′4	47	14	2.5	$Q^{21},K^{7},G^{6},T^{4},E^{3},H^{3},N^{1},R^{1},M^{1}$	**C″ D1**	37	14	2.3	$D^{14},P^{13},E^{9},Q^{4},S^{2},H^{2},A^{2},R^{1}$
CC′5	46	54	3.0	$S^{25},A^{12},T^{4},P^{3},L^{3}$	**C″ D2**	12	5	4.4	$S^{20},K^{18},D^{3},H^{2},A^{2},T^{1},E^{1}$
CC′6	47	95	1.7	P^{39},F^{3},I^{3},V^{2}	**C″ D4**	4	0	7.0	$K^{33},Q^{8},R^{4},I^{1},T^{1}$
C′5	43	4	1.0	$K^{29},R^{10},Q^{4},T^{3},V^{1}$	**F5**	8	3	0.7	$V^{25},T^{8},M^{5},I^{4},L^{4},R^{1}$
C′6	47	63	1.0	$L^{34},V^{3},G^{3},P^{2},R^{2},F^{1},T^{1},M^{1}$	**F7**	47	41	0.5	Y^{36},F^{11}
C′7	1	0	1.2	L^{43},W^{2},V^{1},M^{1}	**F9**	3	0	0.6	$A^{31},V^{5},T^{5},S^{2},N^{2},G^{1},D^{1}$
C′8	3	0	4.0	I^{44},V^{3}	**F10**	1	0	1.6	$R^{32},G^{6},T^{2},A^{2},S^{2},V^{1},Q^{1},N^{1}$
C′9	44	50	1.0[e]	$Y^{34},F^{4},K^{4},G^{3},H^{1},S^{1}$	**FG3**	1	0	4.6	E^{2},D^{1}
C′10	34	23	10.5	$K^{10},Y^{7},G^{5},D^{4},N^{4},A^{3},W^{3},E^{2},F^{2},R^{2},S^{1},H^{1},Q^{1}T^{1},L^{1}$	**FG6**	3	4	5.3	$V^{2},G^{1},W^{1},H^{1},A^{1},D^{1},M^{1}$
C″3	14	4	1.1	$N^{17},T^{9},K^{6},S^{5},Q^{4},R^{3},M^{1},E^{1}$	**FG7**	3	2	5.9	$G^{2},P^{2},I^{1},S^{1},D^{1},L^{1}$
C″4	3	1	1.3	$R^{23},L^{17},S^{5},Q^{1}$	**FG8**	4	3	14.6	$F^{3},E^{2},D^{2},S^{2},P^{1},A^{1},H^{1},Y^{1},I^{1}$
C″5	37	33	1.7	$A^{13},F^{10},E^{6},I^{4},P^{3},Q^{2},D^{2},H^{2},G^{1},R^{1},Y^{1},V^{1}$	**FG9**	6	5	15.0	$G^{5},P^{4},R^{3},C^{1},S^{1},D^{1},E^{1},Q^{1},I^{1},Y^{1}$
C″ D1	19	10	2.9	$S^{31},T^{6},P^{4},D^{3},E^{2}$	**FG10**	9	3	16.5	$D^{5},Y^{4},E^{3},G^{3},H^{3},A^{2},S^{1},N^{1},V^{1},P^{1},Q^{1},T^{1}$
F5	16	5	0.6	$V^{18},T^{9},I^{7},D^{6},M^{3},S^{2},R^{1},N^{1}$	**FG11**	10	7	14.1	$G^{6},Y^{6},D^{4},L^{2},Q^{2},S^{2},T^{2},E^{2},V^{2},N^{1},I^{1},P^{1},R^{1}$
F7	47	67	0.7	Y^{29},F^{18}	**FG12**	8	5	10.3	$R^{8},Y^{6},G^{5},A^{3},T^{3},L^{3},S^{2},H^{2},W^{1},N^{1},P^{1},D^{1},V^{1}$
F9	42	20	0.6	$Q^{25},S^{5},F^{5},A^{5},L^{4},H^{2},V^{1}$	**FG13**	29	25	9.8	$Y^{7},G^{6},D^{5},T^{5},S^{4},L^{3},E^{2},I^{2},N^{2},H^{1},P^{1},W^{1},R^{1}$

Table 1 (*continued*)

VL domain					VH domain				
Position[a]	**Freq.**[b]	**Contact surface (Å²)**[c]	**Cluster radius (Å)**[d]	**Amino acid composition**	**Position**[a]	**Freq.**[b]	**Contact surface (Å²)**[c]	**Cluster radius (Å)**[d]	**Amino acid composition**
F10	1	0	1.2	Q^{33},N^{3},H^{3},L^{3},S^{2},A^{2},K^{1}	**FG14**	35	53	10.0	Y^{12},D^{6},S^{4},A^{4},G^{4},R^{2},V^{2},T^{1},E^{1},N^{1},F^{1},W^{1},P^{1},C^{1},K^{1}
FG10	37	42	3.0	Y^{12},S^{11},W^{7},G^{7},F^{3},H^{3},N^{2},D^{1},T^{1}	**FG15**	42	60	10.5	Y^{16},G^{8},A^{4},R^{3},H^{3},S^{2},W^{2},D^{2},F^{2},E^{1},N^{1},T^{1}
FG11	13	3	4.4	T^{9},N^{7},Y^{6},S^{5},D^{5},G^{4},W^{3},H^{2},E^{1},C^{1},A^{1},I^{1},K^{1},L^{1}	**FG16**	46	76	9.0	G^{10},A^{8},Y^{7},D^{4},N^{3},S^{2},V^{2},P^{2},R^{2},Q^{1},F^{1},W^{1},T^{1},E^{1},K^{1}
FG12	12	6	5.2	S^{14},H^{12},E^{5},N^{4},R^{3},T^{2},G^{2},Y^{2},A^{1},V^{1},K^{1}	**FG17**	45	92	6.5	F^{21},M^{11},G^{5},L^{3},I^{2},S^{2},N^{1},P^{1},W^{1}
FG13	3	3	4.7	S^{2},V^{1},W^{1}	**G4**	47	48	2.0	D^{32},A^{10},V^{2},N^{1},P^{1},K^{1}
FG14	2	0	3.9	P^{1},L^{1}	**G5**	22	12	1.4	Y^{35},V^{5},H^{3},I^{2},F^{1},C^{1}
FG15	44	63	5.5	V^{7},N^{6},S^{5},L^{5},P^{4},F^{4},W^{3},Y^{3},T^{3},R^{2},D^{1},A^{1}	**G6**	47	111	1.1	W^{47}
FG16	46	52	5.3	P^{34},A^{3},L^{3},F^{2},H^{2},R^{1},S^{1},Y^{1}	**G7**	47	22	1.0	G^{47}
FG17	47	109	2.6	L^{11},W^{10},Y^{8},R^{6},P^{5},F^{3},I^{2},A^{1},S^{1}	**G8**	47	15	1.3	Q^{39},A^{5},H^{1},P^{1},E^{1}
G5	30	2	1.6	T^{40},V^{4},I^{2},R^{1}	**G9**	1	0	1.3	G^{47}
G6	47	127	1.3	F^{47}					
G7	47	3	0.8	G^{47}					
G8	47	15	1.1	G^{30},A^{8},Q^{4},S^{4},T^{1}					
G9	3	0	0.7	G^{47}					
G′2	8	2	0.8	K^{43},R^{4}					

[a] Designation of residue positions is based on secondary structure alignment as described.[9] The rationale for FG loop designations is given in Materials and Methods.
[b] The number of entries in which the given position appears at the interface.
[c] Interface contact surface area for a given position averaged over 47 structures. An average contact surface equal to zero means that this position is present at the interface only in a few cases and dividing by 47 gives a value <0.5 Å^2.
[d] Cluster radius, determined following superimposition of all 47 structures as the distance from the center of the cluster of C^{α} atoms for a given position to the most distant atom.
[e] The position of the atom from structure 7fab is anomalous (by more than 6 standard deviations from the average) and was not included in the calculation.

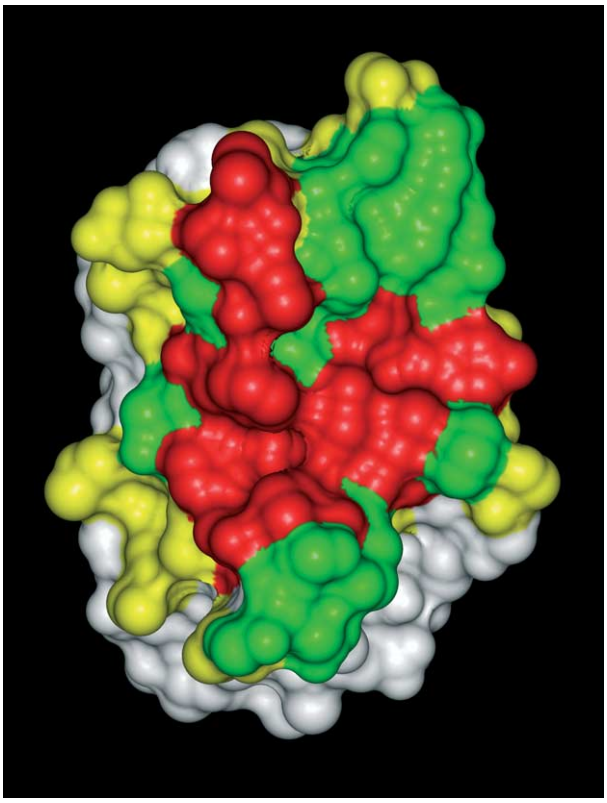

Figure 3. Pictographic relationship of the Ig interface and interface core. A model of the VL domain of PDB file 1a4j (Diels-Alderase Antibody) is shown in a cut away view with the VH domain removed. Regions represented are: VL domain = white + yellow + green + red; virtual interface = yellow + green + red; interface = green + red; interface core = red. Entire residues are colored in all cases. We note that although the VL interface core is composed of the same structural positions for all members of the data set, the red region can vary in shape from structure to structure due to side-chain flexibilities and amino acid compositional heterogeneity (Table 1) at a given core position. The figure was created with Connolly presentation[54] using InsightII software (Accelyris Inc.).

interface (yellow + green + red), the interface formed by the contacting residues in a particular structure (green + red) and the VL–VH interface core (red). The reduced number of contact positions and contact area of the core is evident. We calculate the contact area of the interface core to be 622 Å^2, which represents about 58% of the average interface. Most VL–VH core contacts are hydrophobic with a surface area of 494 Å^2. The remaining contacts are almost invariably hydrophilic and form hydrogen bonds. A few hydrophobic positions (C′6L, C′9L and FG17L) are infrequently occupied by hydrophilic residues (Table 1). However, in these instances as well, hydrogen bonds are formed either with a different partner or the backbone of the original partner. Our list of positions forming the VL–VH interface core extends and refines a list of 12 conserved interface positions described earlier.[46] While these authors included position F9 from VH, the low frequency of this position (Table 1) excludes it from our list. Conversely, L domain positions C5, C′6, C′9, FG16, FG17, and H domain position G4, were not resolved by Chothia *et al*.[46]

A superimposition of all 47 VL domains is presented in Figure 4. The eight interlock positions[9] important for fold formation and common to all sandwich-like proteins were used as reference points. Clusters formed by the interface core positions are seen to be quite compact. For strand positions, the maximal distance of a cluster member from its cluster center is <2.0 Å (Table 1). A single exception is in file 7fab (Human immunoglobulin fragment), where an unusual organization of the C′, C″ region results in a large dispersion at position C′9 of VL domains. In addition, there are three interface core positions in the FG loop with large contact surface areas that are dispersed.

CL–CH1 interface

The CL–CH1 interface is formed predominantly of Sheet I residues from these domains. Fifty CL and 52 CH1 positions compose the virtual interface (Table S1 in Supplementary Material). For a given structure, 24 to 38 residues (average, 32) from CL and 23 to 35 residues (average, 29) from CH1 make up the interface, yielding an average interface contact surface of 1233 Å^2 (maximum, 1443 Å^2; minimum, 891 Å^2). Thus, while the VL–VH and CL–CH1 virtual interfaces are composed of approximately the same number of residues, the average number of residues and surface area buried upon complex formation is larger for CL–CH1. This might mean that a larger number of random residue replacements occurred during evolution of the variable domains, making them less complementary. Indeed, a comparison of the data reveals that residues at the CL–CH1 interface are 2 to 3 times more conserved than those at the VL–VH interface (compare Table 1 with Table S1 in Supplementary Data).

The CL–CH1 interface core is composed of 18 primary positions from CL (A7, A9, AB1, AB3, AB4, AB7, B5, B7, B9, B11, B12, D7, D9, D14, E3, E5, E7, E9), and 14 from CH1 (A7, A9, A10, B5, B9, B11, D7, D9, D10, D12, D14, E5, E7, G7). This core was obtained (as described above for VL–VH) from residue–residue contacts forming the "80% core" for CL–CH1 by eliminating contacts that are either non-attractive, infrequent or heterogeneous in residue type (see Figure 1S in Supplementary Data). The CL–CH1 interface core area amounted to 691 Å^2 (314 Å^2 hydrophobic and 377 Å^2 hydrophilic), which represents 56% of the average interface.

The above list of 32 residues forming the CL–CH1 interface core refines the set of residues found by Miller[47] employing four complexes. Miller's list included positions B10 and D12 from the CL domain, and positions B6 and E4 from CH1.

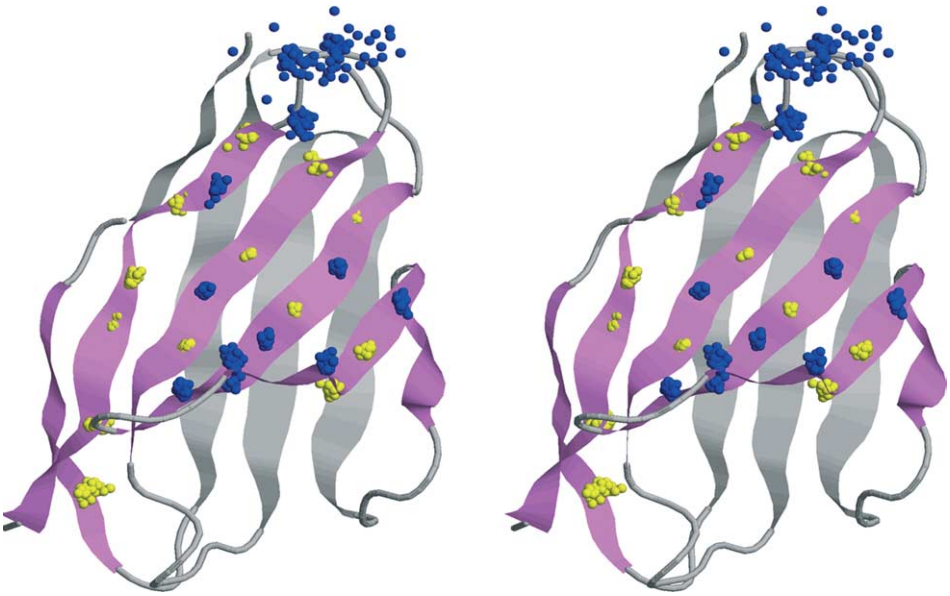

Figure 4. Spatial relations of the domain and interface core positions. Superimposition of the interface and domain core positions for all 47 VL domains was performed using the universal sandwich-like proteins interlock positions[9] as reference points. Strands of the β-sheet taking part in domain-domain contact are colored pink. Clusters colored yellow show the C^{α} positions for the domain core, while those colored blue show the C^{α} positions for the interface core. Note that most blue cluster positions are immediately adjacent to yellow ones. The gentle undulation of the strands, apparent in the stereo view, is a frequent feature of β-sheet geometry. Correlatively, side-chains of residues at peak and valley positions head in opposite directions. This produces staggered core positions and encourages the neighboring seen for interface and domain core positions. The structure of PDB file 12e8 (Apolipoprotein-E Fab fragment) was used to draw the ribbon-wire frame model, employing Rasmol software.

However, infrequency at the interface, or absence from the minimal set of residues forming the "80% core", indicates that they are not primary positions. Furthermore, while Miller[47] included position A8 from CH1, we do not because mostly non-attractive contacts are formed at this position. On the contrary, we found residues at position B9, B12 and E9 of the CL domain, and A10 of the CH1 domain, to be part of the interface core.

In summary, the average total interface between the L and H chains is about 2300 Å^2, which is within the size range of interfaces yielding strong binding.[12] Both VL–VH and CL–CH1 experience about a halving in number of residue positions going from virtual to average interface and then again from average to core interface. It is tempting to speculate that the reduced set of positions making up the interface core is responsible for initial protein–protein recognition and/or nuclear formation. It may be that properties of the cores differ from those of the interfaces as a whole. If so, schemes for the recognition of such surface patches might be more efficient.

Domain cores

The procedure described for obtaining the interface core was used to identify the domain cores, except that only positions forming sheet–sheet contacts were considered. Positions at which residues form intra-domain sheet–sheet contacts for VL, VH, CL and CH1, and the minimal sets of residue–residue contacts forming the "80% core" of the domain interfaces, can be found in Tables S2–S5 and Figures S2–S5 in Supplementary Data. The residues at the domain core positions are the main

Table 2. Characteristics of the internal sheet–sheet interfaces within Ig domains

Domain	Number of residues in sheets I and II						Contact surface area (Å^2)			
	Virtual interface		Sheet–sheet interface[a]		Domain core		Sheet–sheet interface[b]		Domain core	
	I	II	I	II	I	II	Full	Phobic[c]	Full	Phobic[c]
VL	26	29	19	21	10	13	922	607	629	471
VH	24	31	20	23	11	14	1008	723	711	549
CL	22	20	18	12	12	10	768	661	617	540
CH1	25	20	18	13	12	9	695	558	510	449

[a] The number of sheet–sheet interface positions averaged over 47 structures.
[b] The sheet–sheet interface surface area averaged over 47 structures.
[c] The hydrophobic component of the contact surface area.

effectors of interaction between the two β-sheets within the VL, VH, CL and CH1 domains.

The characteristics of the domain interfaces and cores are summarized in Table 2. The relatively modest decrease in number of residues from virtual to average interface (27%) and from average interface to domain core (36%) for all four domains, points to a degree of uniformity among Ig proteins at the domain fold level. We note that this uniformity holds as well at the secondary structure level, with 42% of domain cores occupying the same structural positions in VL, VH, CL and CH1 (cf. Figure 5(a) and (b)). As can be derived from Table 2, the intra-domain interfaces, and, more so, the domain cores, are highly hydrophobic (average of 76% and 82%, respectively), while the non-core areas are less so (average of 61%). Superimposition of 47 structures yielded a small dispersion (<2.0 Å) of the domain core C^{α} atom positions (Tables S2 and S5 in Supplementary Data).

Note that the calculation of the domain core in this research is based on principles different from those employed to determine the core structure described by Gershtein & Altman.[55] The latter characterized their core as a subset of atoms with low structural variation and included residues that are surface located. We define the domain core as a set of residue pairs that form the main contacts between two sheets within a domain. These residues are almost invariably buried. The domain core positions shown in Figure 6 are a subset of the Ig core structure positions described by Gelfand *et al.*,[56,57] and include the interlock positions common to all sandwich-like proteins.[9]

Relationship between interface cores and domain cores

The structural sequence alignment of the VL, VH, CL and CH1 domains presented in Figure 5 highlights positions of the interface core in blue and the domain core in yellow. We find that most (66%) of the residues forming the interface core are immediately adjacent (i.e., covalently linked) to residues forming the domain core (also see Figure 4). In all 47 structures, the interlocked strands (viz., C and F, and B and E in the variable and constant domains, respectively) are represented at the interface core. Interlocked strands are common to virtually all sandwich-like proteins.[9] Within these strands in our Ig data set, the interlock positions (viz., B6, B8, C4, C6, E8, E10, F6, F8) are immediately adjacent to interface core positions 65% of the time in the variable domains and 80% in the constant domains. Non-interlocked β strands (viz., C′ and G, and A and D in the variable and constant domains, respectively) likewise contribute conserved residues adjoining the interface and domain cores

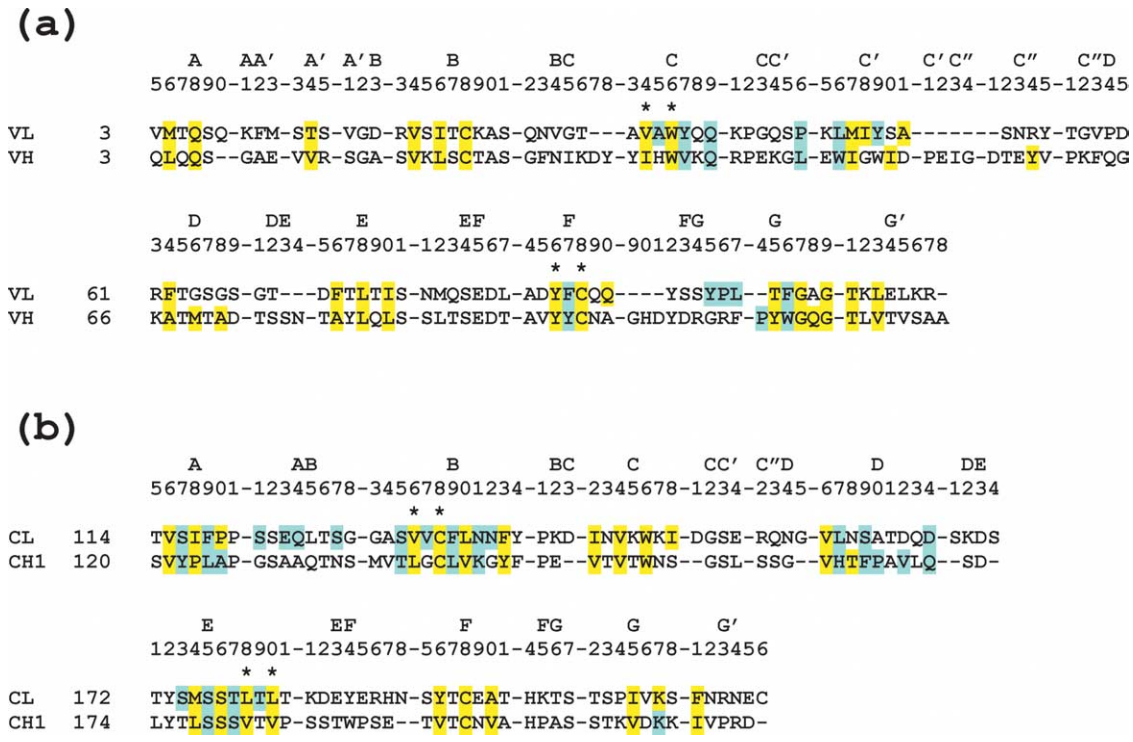

Figure 5. Structural sequence alignment of Ig domains. Secondary structure assignment was done as described by Kister *et al*.[9] The sequences shown are for PDB entry 12e8. For clarity, a gap was inserted between strands and loops. Interlock positions of the interface sheets are starred. Domain core positions are colored in yellow and interface core positions are colored in blue. Position G7 of CH1, which is located in the strand at the edge of the sheet, is colored blue since in virtually all cases it forms a strong hydrogen bond with a loop of the CL domain. However, it also contributes to formation of the hydrophobic domain core by its side-chain aliphatic part. (a) VL and VH domains. (b) CL and CH1 domains.

PDB entry		Strand *i*		Strand *k*	
ID	Chains	Initial Position	Sequence	Initial position	Sequence
			* *		* *
1a3q	A:227-327	244	-DEVYLLCD-	285	-YAIVFRTP-
1adw	A	28	-GDVINFVPT-	63	-SYTLTVT-
1aoh	A	36	-ANCDFVYSY-	78	-KMIVFLFAE-
1bfs	A	264	-EEIYLLCD-	307	-FAIVFKTP-
1c16	A:181-276	197	-GDVTLRCWALGFY-	240	-TFQKWAAVVVP-
1cd8	A	32	-GCSWLFQP-	89	-EGYYFCSALSN-
1dqi	A	48	-HIRYIELYFLPE-	102	-KGKLYALSYCN-
1dqt	A	16	-VASFPCEYSP-	73	-SRVNLTIQ-
1exu	A:177-267	192	-GFSVLTCSAFSFY-	233	-SFHASSSLTVKS-
1fat	A	64	-TVASFATSFTFN-	219	-ETNDVLSWSFASKL-
1f5w	A	52	-PLDIEWLISPA-	115	-IGTYQCKVKKA-
1gzc	A	68	-ASFETRFSFSI-	225	-THDVYSWSFQASLP-
1gzq	A:184-280	200	-GRLQLVCHVSGFY-	242	-TWYLRATLDVAD-
1gzt	A	12	-TRFGVTAFAN-	50	-IGTQVLNS-
1ic1	A:1-82	16	-SVLVTCST-	48	-NRKVYELS-
1imh	C:368-468	384	-EEVFLIGKN-	426	-NHLIVKVP-
1k5n	A:182-276	197	-HEATLRCWALGFY-	240	-TFQKWAAVVVPS-
1kgc(v)	E:3-118	30	-VSLFWYQQA-	87	-SAVYLCASSL-
1kgc(c)	E:119-247	142	-QKATLVCLATGFY-	190	-RYCLSSRLRVSA-
1my7	A	210	-DEIFLLCD-	248	-VAIVFRTP-
1nls	A	102	-ETNTIL-	188	-VASFEATFTFL-
1spp	A	40	-YKLLVSIPTLNL-	102	-PYEIIFLRDS-
2bb2	A:86-175	130	-TWVGYQYP-	164	-VQSVRRI-
3fru	A:179-269	194	-GSSVLTCAAFSFY-	235	-SFHAWSLLEVKR-

Figure 6. Partial structural sequence alignment of non-Ig sandwich-like domains. The *i* and *k* interlock strands[9] from the interface sheet of the 24 PDB structures in the non-Ig data set were aligned by superimposing their interlock positions (starred). Residues included in formation of the "80% domain core" are colored yellow, while those included in formation of the "80% interface core" are colored blue. Chain E in PDB entry 1kgc has two sandwich-like domains: variable (v) and constant (c).

(Figure 5). These core positions are important for Ig fold and Ig interface stabilities but are not necessarily present in non-Ig sandwich-like proteins.

Non-Ig sandwich-like proteins

The immunoglobulin superfamily represents an often crystallized but relatively small subset of sandwich-like proteins. The latter encompass 82 superfamilies and 38 protein folds in the SCOP database.[58] We applied our procedures for Ig proteins to analyze domain interactions in other sandwich-like complexes coupled in a sheet–sheet mode. A non-redundant, high-resolution set of 24 structures was extracted from the PDB as described in Materials and Methods. The interface characteristics between and within domains for these structures are presented in Table 3, while the averages for the Ig and non-Ig structures are compared at the bottom of the table. Not surprisingly, the domain lengths of the contacting sheets for the non-Ig set are, more heterogeneous than those for the Ig set, and, associatively, so too are the domain "80% cores". However, the average interface areas, and ranges (604 Å^2 to 1683 Å^2 for non-Ig and 800 Å^2 to 1443 Å^2 for Ig), are similar for the two sets, as are the average number of residues, and ranges (9 to 34 for non-Ig and 11 to 30 for Ig), of the interface "80% core".

Many of the proteins in our non-Ig data set (all except 1f5w, 1cd8, 1kgc, 1dqt) seem to have a flatter interface than the barrel-like VL–VH interface that is highly curved. Presumably, more orientations between domains are possible for flatter surfaces than for curved ones. In this regard, it is relatively easy to dock VL–VH interfaces with rigid body docking software, but less so other sheet proteins, even when the binding site is known.

The data summarized in Table 3 show that, for most cases, a majority of the non-Ig interface "80% core" residues adjoin those of the domain "80% core". The level of juxtaposition (57% on average) is in fact higher than for the "80% core" in the Ig data set (37% on average). This is fortunate, for while the Ig data set can be statistically analyzed to arrive at refined domain and interface core structures, the members of the non-Ig set must be treated as individual structures due to substantially different strand topologies. Nonetheless, a partial alignment of non-Ig structures was achieved by first identifying the interlock strands, and within them the interlock positions, as described by Kister *et al.*[9] The *i* and *k* strands were then aligned by superimposition of the interlock positions, as shown in Figure 6. Clearly, in a majority of cases, interlock and interface "80% core" positions are found one next to the other. Moreover, neighboring of domain and interface "80% core" positions occurs along the lengths of most of the interlock strands (Figure 6).

The tendency for neighboring carries over to the non-interlock strands as well. For example, the non-Ig data set contains three structures (1gzc [*Erythrina cristagalli* lectin], 1nls [concanavalin A], 1fat [phytohemagglutinin-L]), where the interface "80% core" between the interacting domains is formed by a relatively small component of two

Table 3. Characteristics of intra- and inter-domain interfaces in non-Ig proteins

PDB ID	Chains[a]		Interface area[b] (Å^2)	Number of residues									
				Domain lengths[c]		Domain "80% core"[d]		Interface "80% core"[e]		Interface "80% core" juxtaposed with			
										Domain "80% core"		Interlock positions[f]	
	1	2		1	2	1	2	1	2	1	2	1	2
1a3q	A:227–327	B:227–327	938	101	101	24	19	13	12	7	5	3	3
1adw	A	X	613	123	123	21	29	11	14	7	9	4	4
1aoh	A	B	805	143	147	34	33	14	15	10	10	4	3
1bfs	A	X	956	106	106	14	14	11	12	6	7	3	4
1c16	A:181–276	B	728	96	99	28	27	13	9	11	5	4	1
1cd8	A	X	1330	114	114	26	26	22	24	10	11	3	4
1dqi	A	B	1571	124	124	25	24	25	25	7	8	4	5
1dqt	A	B	907	117	117	30	32	17	17	14	14	6	6
1exu	A:177–267	B	974	91	99	29	27	15	13	11	8	4	4
1f5w	A	B	939	124	121	29	27	15	16	9	8	2	2
1fat	A	C	747	232	232	45	44	13	14	11	11	0	0
1gzc	A	B	1008	239	239	44	44	16	15	12	12	1	1
1gzq	A:184–280	B	893	97	100	20	17	15	10	7	6	3	2
1gzt	A	B	1684	114	114	33	30	31	34	15	16	5	5
1ic1	A:1–82	X:1–82	604	82	82	34	34	11	9	9	7	4	3
1imh	C:368–468	D:368–468	758	101	101	16	16	10	11	4	5	3	4
1k5n	A:182–276	B	780	95	100	16	19	13	11	5	6	4	3
1kgc	D:118–206	E:119–247	1678	89	129	24	21	27	28	13	12	5	6
1kgc	D:2–117	E:3–118	887	112	112	25	26	13	15	3	4	1	1
1my7	A	B	902	107	102	16	15	12	14	7	8	3	3
1nls	A	B	1395	237	237	45	45	25	27	13	14	0	1
1spp	A	B	959	109	112	21	24	17	15	11	9	4	6
2bb2	A:86–175	X:2–85	1127	91	87	13	10	19	20	7	6	3	4
3fru	A:179–269	B	854	91	99	17	18	13	13	8	8	4	4
Non-Ig (averages)			1002	119		26		16		9		3.2	
Ig (averages)			1154	108		19		19		7		3.3	

[a] The full chain makes up the β sandwich-like domain unless otherwise indicated by residue numbering. Chains marked as X were obtained by applying crystal symmetry operations.
[b] The interface area was calculated as the sum of the atom–atom contacts, as defined by CSU software.[50]
[c] As counted in the coordinate section of the PDB entry.
[d] Arrived at by taking the minimal number of pair positions forming 80% of the intra-domain interface.
[e] Arrived at by taking the minimal number of pair positions forming 80% of the inter-domain interface.
[f] As defined in Kister *et al*.[9]

Table 4. Amino acid conservation at Ig interfaces

		P values					
		Valdar & Thornton[35]			Mirny & Shakhnovich[60]		
Ig domain	Number of homologues	Interface	80% core	Core	Interface	80% core	Core
VL	405	0.84	0.57	0.25	0.85	0.73	0.18
VH	987	0.71	0.76	0.04	0.56	0.75	0.08
CL	86	0.05	0.04	0.08	0.04	0.03	0.08
CH	220	0.00	0.09	0.01	0.00	0.00	0.00

The domain sequences of PDB entry 12e8 were used to derive the HSSP homolog families. The *P* values for compositional conservation at the interfaces were determined *versus* all solvent exposed residues of the domain. *P* values <0.005 are rounded to 0.00.

large β sheets, each with more than 100 residues and containing six strands. In structures 1gzc and 1nls, only two of the four interlock strands form part of the interface "80% core" while in 1fat none of them do. (The interlock strands for chain A of each of these cases are shown in Figure 6.) However, in these three structures the overall percentage of interface "80% core" positions neighboring to domain "80% core" positions is quite high (70% on average, as derived from Table 3). Thus, in these cases, the proposed rigidity at the interface provided by neighboring must be provided mainly by interactions involving the non-interlock strands. Based on these and several other cases, we speculate that the interlock strands, and within them the interlock positions, operate at a formative, nucleation stage in the sheet–sheet interaction of sandwich-like proteins, while rigidity per se, afforded by the juxtaposition of domain and interface core positions, is a general feature of the core structures that is shared by interlock and non-interlock strands alike. In summary, for sandwich-like proteins interacting in a sheet–sheet mode the unique relationship described above between domain and interface cores is likely to be involved in anchoring the protein recognition site, increasing local rigidity, and explaining rapid binding.

Amino acid conservation at the non-Ig interface

Can the resolved non-Ig structures described in Table 3 serve as a template for predicting sandwich-like interfaces of unresolved homologous proteins? As a criterion, we chose amino acid positional conservation. We searched for an increase in compositional conservation, progressing in succession from exposed, to interface, to interface core residues among sets of homologous sequences. The HSSP database was used to provide a multiple sequence alignment of homologues and a sequence profile characteristic of the protein family, centered on a known structure.[59]

The relevance of the approach to our case was first tested using Ig proteins. We determined the occurrence of individual amino acid residues, or common groupings of amino acid residues, at interface positions and calculated the level of compositional conservation of these positions against all solvent exposed positions. The statistical significance (*P* values) of the conservation was determined by two methods[35,60] and is summarized in Table 4. For constant domains, the results indicate that the full interface is already highly conserved *versus* all exposed residues ($P \leq 0.05$). For variable domains, the results suggest a progression to greater amino acid conservation going from interface to core, however the significance of the increase is statistically clear only occasionally. An explanation might be that only 12 of the 20 positions in the variable interface are strongly conserved compositionally.[46]

The results were sufficiently encouraging to warrant examination of the non-Ig proteins

Table 5. Amino acid conservation at the "80% core" of non-Ig proteins

		P values	
Known PDB structure	Number of homologues	Valdar & Thornton[35]	Mirny & Shakhnovich[60]
1f5w	49	0.67	0.11
1cd8	108	0.19	0.05*
1kgc(v)	243	0.07	0.08
1dqt	35	0.56	0.45
3fru	380	0.01*	0.02*
1k5n	1253	0.00*	0.00*
1c16	448	0.00*	0.00*
1exu	139	0.01*	0.02*
1kgc(c)	61	0.01*	0.09
1gzq	61	0.15	0.19
1ic1	33	0.94	0.75
1bfs	78	0.01*	0.00*
1a3q	80	0.00*	0.00*
1my7	66	0.02*	0.00*
1dqi	23	0.59	0.02*
1aoh	11	0.85	0.74
1adw	59	0.24	0.84
2bb2	127	0.42	0.01*
1spp	32	1.00	0.16
1nls	9	0.59	0.57
1gzc	209	0.31	0.48
1fat	211	0.15	0.85
1imh	28	0.28	0.19
1gzt	5	0.04*	0.09

Statistically significant conservations is taken as $P \leq 0.05$. *P* values <0.005 are rounded to 0.00.

described in Table 3. We analyzed the amino acid conservation of the interface and "80% core" residues against the set of exposed domain residues for each of the 24 homologous families of proteins. The P values for the interface and "80% core" were generally similar and only the latter are given in Table 5. Approximately 40% of the cases showed highly significant P values. These include the MHC-related protein families (3fru, 1k5n, 1c16, 1exu) and the NF-κB transcription factor families (1bfs, 1a3q, 1my7). The arch type proteins in these cases have a large number of close homologs (30% to 100% sequence identity) having the same oligomeric state and similar functions. On the other hand, only small increments in amino acid conservation of the "80% core" *versus* the total complement of exposed residues was found for the tetrameric lectin families (1nls, 1gzc, 1fat) and several poorly populated families (e.g., 1dqt, 1ic1, 1aoh, 1adw, 1spp).

Figure 7 graphically presents amino acid conservation as a function of sequence position for HSSP homolog families 1c16, 1dqt and 1kgc (variable domain). The horizontal red line indicates the average, exposed-amino acid conservation level for a given set of homologous proteins. One can distinguish cases with highly significant P values (1c16 set, $P=0.001$), where most interface points are atypical and distributed well above the red line, from those with obviously non-significant values (1dqt set, $P=0.45$), where the distribution of interface points are typical of the distribution of all the points as a whole. For the former, the high level of sequence homology at the "80% core" positions indicates that this domain has an equivalent function (i.e., complex formation) and is strictly structured (limited evolution) throughout the 448-member 1c16 homolog set. However, for the 35 homologs of the 1dqt set, functional predictions cannot be made based on the conservation distribution pattern.

A third case is represented by the 1kgc(v) homolog set (Figure 7). Conservation of the "80% core" positions varies across the sequence profile and, on average, is not significantly different than for exposed residues in general ($P=0.08$). However, when the profile is analyzed segmentally, a stretch of exposed positions (region A) that includes 7 positions of the "80% core" and is highly conserved over background ($P<10^{-3}$) can be recognized. So too, another stretch (region B) which includes 4 positions of the "80% core" but is one of the least conserved regions of the profile ($P=0.99$). Recognition of such segments can be of value in planning or understanding mutational modifications that may affect complex formation.

The positional conservation of interface residues for the remaining 21 non-Ig protein homolog sets is given in Figure S6 of Supplementary Material. For a majority of the sets, statistically relevant information can be extracted from the profiles to analyze sandwich-like interfaces.

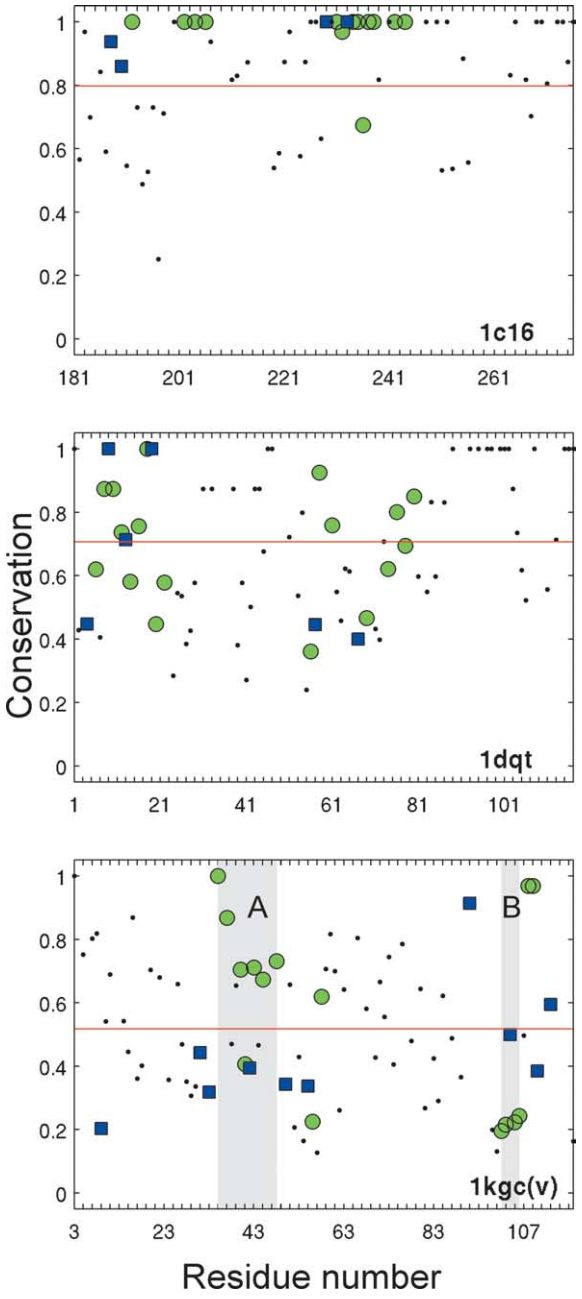

Figure 7. Amino acid conservation as a function of sequence position. The labels in the lower right corners identify the homolog sets. Conservation was calculated using the approach of Mirny & Shakhnovich.[60] Only solvent exposed positions are shown. Residue numbering corresponds to that for PDB entries 1c16, 1dqt and 1kgc (variable domain). Green circles mark "80% core" residue positions, blue squares mark the additional positions that complete the interface while black dots mark the remaining positions of the exposed residues. The horizontal red line indicates the mean level of conservation for all solvent exposed positions.

Conclusions

We have demonstrated that where statistical data exist (as in the case of the resolved heavy and light chains of Ig proteins), subsets of interface positions

can be extracted that define an interface core and a domain core, and offer a new tool with a reduced target for the analysis of sheet–sheet interactions in sandwich-like proteins. The refined interface cores contain approximately half the residues and half the surface areas of the full interfaces. The techniques developed for the Ig interfaces and domain cores proved adequate to extract first-approximation cores for a set of individual non-Ig sandwich-like interfaces, thus extending the usefulness of the approach.

Our analysis revealed that most of the positions in sandwich-like proteins crucial for sheet–sheet inter-domain and intra-domain interactions are adjoined one to another in sequence. This finding was derived independent of geometric considerations, however β-sheet side-chain geometry clearly dictates such neighboring. Since the domain core is commonly accepted as the most stable part of the protein structure, we can expect that adjoining residues will also be rather restricted in their flexibility. The tight clustering of the interface core positions across the Ig data set bears this out. Thus, the juxtaposition of interface and domain core residues documented here experimentally supports the concept of a rigid substructure on the protein surface involved in complex formation.

Is such neighboring a feature of domain–domain interactions for structural elements other than β sandwich sheets? We currently have no answer for this. A logical next step will be to look at β-sheet to non-β sheet interfaces. We note that loop regions, which normally have reduced geometric constraints, also participate in protein–protein interface formation. It will be interesting to examine if the concepts described here can be extrapolated to these regions as well.

Materials and Methods

Databases

Ig database. Two hundred and eighty one structures of Fab fragments from the Protein Data Bank (PDB)[42,43] were extracted using structural classification of protein assignments (SCOP v. 1.59).[58] Structures of resolution 2.3 Å or better were retained, and Fab fragments selected such that none had more than 80% sequence identity with any other. The final database of 47 Ig structures is composed of the following PDB entries: 12e8, 1a3l, 1a3r, 1a4j, 1aqk, 1bm3, 1c1e, 1c5c, 1ce1, 1ct8, 1d5i, 1dn0, 1dqq, 1e6o, 1emt, 1f3d, 1f58, 1fe8, 1flr, 1fns, 1g9m, 1hil, 1hyx, 1i8m, 1il1, 1ind, 1iqd, 1jfq, 1jgl, 1jgu, 1jps, 1k4c, 1kel, 1mfe, 1nbv, 1osp, 1qkz, 1sm3, 1vge, 1wej, 1yej, 2fb4, 2fbj, 2hrp, 2pcp, 7fab, 8fab.

Non-Ig database. All β-sandwich domains (6915 domains from 2177 PDB entries) were identified using SCOP (version 1.63).[58] Those predicted by the PQS database[61] to have a non-monomeric state (1410 entries) were retained. From these, all pairs of interacting β-sandwich domains (2959 pairs) were determined using CSU software.[50] For pairs having identical domain IDs (493 pairs), the pair with the best resolution was retained. All file pairs with interfaces having a contact area less than 500 $Å^2$ (232 pairs) were then discarded. A literature-based search for monomeric molecules was carried out among the remaining 261 pairs, leaving 131 proven non-monomeric pairs. In a final culling, unusual cases (such as long chain dimers where the interface of interest is a very minor component, dimers where positional strand swapping occurs within the pair, non-oligomeric complexes or multimeric proteins) were discarded. This resulted in a dataset of 87 pairs of which 24 are examples of sheet–sheet mode interface interaction. These 24 files are listed in Table 3.

Interlock positions of Ig domains

Immunoglobulin molecules are built of two heavy and two light chains. A light chain folds into two domains, designated variable (VL) and constant (CL), while the heavy chain folds into four domains: a variable one (VH) and three constant ones (CH1, CH2 and CH3). The light and heavy chains are held together by disulfide bridges and by association of VL with VH, and CL with CH1, in a sheet–sheet mode.[62] These four domains are each built of two approximately parallel β sheets (Sheet I and Sheet II). The variable and constant domains differ in the number of strands that compose the sheets (Figure 1). Contact between variable domains is formed by Sheets II, while contact between constant domains is formed by Sheets I.

Two interlocked pairs[9] of β strands (B, E and C, F) are located at the center of each domain. Eight conserved, hydrophobic, residue positions (B6, B8, C4, C6, E8, E10, F6 and F8) within the interlocked strand pairs form the interlock positions[9] of the Ig domains. The C^{α} atoms of residues at these eight conserved positions were used as reference points for superimposition of structures. Clusters formed by the interlock positions upon such superimposition are quite compact (the maximal distances from cluster centers were <1 Å), showing that distances between these positions are geometric invariants.[56]

Linear residue positions

The definition of linear residue positions adopted here is based on secondary structure alignment as described.[9] A difference, however, is the designation of positions in the loop between F and G strands (FG loop). The FG loop varied strongly in length from 3 to 17 residues. Our analysis showed that the ultimate loop residue in all cases was structurally homologous (independent of loop length). Thus, in our alignment, the last three positions FG15, FG16 and FG17 are occupied by residues in all structures, while the first positions of the loop may be gapped.

Amino acid conservation of interface positions

Estimation of the conservation for each position in a sequence was done in two steps. First, the HSSP database[59] was used to derive a set of homologous secondary structure protein sequences and their multiple alignments. Then, the approaches suggested by Valdar & Thornton[35] and Mirny & Shakhnovich[60] were applied to determine the conservation of a given position. In the first approach, a mutation data matrix is applied to measure similarity between amino acid residues, while in the second, residues are grouped into classes, ignoring mutation within a class. Since aromatic residues Phe, Tyr and Trp are often part of a hydrophobic core we included them in the aliphatic class for the second approach.[60] Finally, the probability of obtaining the

mean conservation of a subset of solvent exposed residues by chance alone (P value) was calculated as described.[35,60]

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2004.06.072

References

1. Volkenstein, M. V. (1979). Electronic-conformational interactions in biopolymers. *Pure Appl. Chem.* **51**, 801–829.
2. Karplus, M. & McCammon, J. A. (1983). Dynamics of proteins: elements and function. *Ann. Rev. Biochem.* **52**, 263–300.
3. Jimenez, R., Salazar, G., Yin, J., Joo, T. & Romesberg, F. E. (2004). Protein dynamics and the immunological evolution of molecular recognition. *Proc. Natl Acad. Sci. USA*, **101**, 3803–3808.
4. Levinthal, C. (1968). Are there pathways for protein folding? *J. Chim. Phys. Phys.-Chim. Biol.* **65**, 44.
5. Baldwin, R. L. & Rose, G. D. (1999). Is protein folding hierarchic? I. Local structure and peptide folding. *Trends Biochem. Sci.* **24**, 26–33.
6. Berezovsky, I. N. & Trifonov, E. N. (2002). Loop fold structure of proteins: resolution of Levinthal's paradox. *J. Biomol. Struct. Dyn.* **20**, 5–6.
7. Fersht, A. R. (1997). Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.* **7**, 3–9.
8. Finkelstein, A. V. (2002). Cunning simplicity of a hierarchical folding. *J. Biomol. Struct. Dyn.* **20**, 311–313.
9. Kister, A. E., Finkelstein, A. V. & Gelfand, I. M. (2002). Common features in structures and sequences of sandwich-like proteins. *Proc. Natl Acad. Sci. USA*, **99**, 14137–14141.
10. Valerio-Lepiniec, M., Nicaise, M., Adjadj, E., Minard, P. & Desmadril, M. (2002). Key interactions in neocarzinostatin, a protein of the immunoglobulin fold family. *Protein Eng.* **15**, 861–869.
11. Najmanovich, R., Kuttner, J., Sobolev, V. & Edelman, M. (2000). Side-chain flexibility in proteins upon ligand binding. *Proteins: Struct. Funct. Genet.* **39**, 261–268.
12. Lo Conte, L., Chothia, C. & Janin, J. (1999). The atomic structure of protein-protein recognition sites. *J. Mol. Biol.* **285**, 2177–2198.
13. Goh, C.-S., Bogan, A. A., Joachimiak, M., Walther, D. & Cohen, F. E. (2000). Co-evolution of proteins with their interaction partners. *J. Mol. Biol.* **299**, 283–293.
14. Zhou, H. X. & Shan, Y. B. (2001). Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins: Struct. Funct. Genet.* **44**, 336–343.
15. Fariselli, P., Pazos, F., Valencia, A. & Casadio, R. (2002). Prediction of protein–protein interaction sites in heterocomplexes with neural networks. *Eur. J. Biochem.* **269**, 1356–1361.
16. Lichtarge, O. & Sowa, M. E. (2002). Evolutionary predictions of binding surfaces and interactions. *Curr. Opin. Struct. Biol.* **12**, 21–27.
17. Jones, S. & Thornton, J. M. (1997). Prediction of protein–protein interaction sites using patch analysis. *J. Mol. Biol.* **272**, 133–143.
18. Landgraf, R., Xenarios, I. & Eisenberg, D. (2001). Three-dimensoinal cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.* **307**, 1487–1502.
19. Aloy, P., Querol, E., Aviles, F. X. & Sternberg, M. J. E. (2001). Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J. Mol. Biol.* **311**, 395–408.
20. Madabushi, S., Yao, H., Marsh, M., Kristensen, D. M., Philippi, A., Sowa, M. E. & Lichtarge, O. (2002). Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *J. Mol. Biol.* **316**, 139–154.
21. Chothia, C. & Janin, J. (1975). Principles of protein–protein recognition. *Nature*, **256**, 705–708.
22. Argos, P. (1988). An investigation of protein subunit and domain interfaces. *Protein Eng.* **2**, 101–113.
23. Jones, S. & Thornton, J. M. (1997). Analysis of protein–protein interaction sites using surface patches. *J. Mol. Biol.* **272**, 121–132.
24. Larsen, T. A., Olson, A. J. & Goodsell, D. S. (1998). Morphology of protein–protein interfaces. *Structure*, **6**, 421–427.
25. Sheinerman, F. B., Norel, R. & Honig, B. (2000). Electrostatic aspects of protein–protein interactions. *Curr. Opin. Struct. Biol.* **10**, 153–159.
26. Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342–358.
27. Schreiber, G. & Fersht, A. R. (1995). Energetics of protein-protein interactions: analysis of the barnase–barstar interface by single mutations and double mutant cycles. *J. Mol. Biol.* **248**, 478–486.
28. Bogan, A. A. & Thorn, K. S. (1998). Anatomy of hot spots in protein interfaces. *J. Mol. Biol.* **280**, 1–9.
29. Glaser, F., Steinberg, D. M., Vakser, I. A. & Ben-Tal, N. (2001). Residue frequencies and pairing preferences at protein–protein interfaces. *Proteins*, **43**, 89–102.
30. Brinda, K. V., Kannan, N. & Vishveshwara, S. (2002). Analysis of homodimeric protein interfaces by graph-spectral methods. *Protein Eng.* **15**, 265–277.
31. Deitmann, S. & Frommel, C. (2002). Prediction of 3D neighbours of molecular surface patches in proteins by artificial neural networks. *Bioinformatics*, **18**, 167–174.
32. DeLano, W. L. (2002). Unraveling hot spots in binding interfaces: progress and challenges. *Curr. Opin. Struct. Biol.* **12**, 14–20.
33. Kortemme, T. & Baker, D. (2002). A simple physical model for binding energy hot spots in protein–protein complexes. *Proc. Natl Acad. Sci. USA*, **99**, 14116–14121.
34. Ma, B. Y., Elkayam, T., Wolfson, H. & Nussinov, R. (2003). Protein–protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc. Natl Acad. Sci. USA*, **100**, 5772–5777.
35. Valdar, W. S. J. & Thornton, J. M. (2001). Protein–protein interfaces: analysis of amino acid conservation in homodimers. *Proteins: Struct. Funct. Genet.* **42**, 108–124.
36. Fernandez, A., Scott, L. R. & Scheraga, H. A. (2003). Amino acid residues at protein–protein interfaces: why is propensity so different from relative abundance? *J. Phys. Chem. B*, **107**, 9929–9932.
37. Cole, C. & Warwicker, J. (2002). Side-chain conformational entropy at protein–protein interfaces. *Protein Sci.* **11**, 2860–2870.

38. Ofran, Y. & Rost, B. (2003). Analyzing six types of protein–protein interfaces. *J. Mol. Biol.* **325**, 377–387.
39. Aloy, P. & Russell, R. B. (2002). Interrogating protein interaction networks through structural biology. *Proc. Natl Acad. Sci. USA*, **99**, 5896–5901.
40. Selzer, T. & Schreiber, G. (2001). New insights into the mechanism of protein–protein association. *Proteins: Struct. Funct. Genet.* **45**, 190–198.
41. Chakrabarti, P. & Janin, J. (2002). Dissecting protein–protein recognition sites. *Proteins: Struct. Funct. Genet.* **47**, 334–343.
42. Fernandez, A. & Scheraga, H. A. (2003). Insufficiently dehydrated hydrogen bonds as determinants of protein interactions. *Proc. Natl Acad. Sci. USA*, **100**, 113–118.
43. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D. Rodgers, J. R. *et al.* (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.
44. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N. Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucl. Acid Res.* **28**, 235–242.
45. Chothia, C. & Janin, J. (1981). Relative orientation of close packed beta-pleated sheets in proteins. *Proc. Natl Acad. Sci. USA*, **78**, 4146–4150.
46. Chothia, C., Novotny, J., Bruccoleri, R. & Karplus, M. (1985). Domain association in immunoglobulin molecules: the packing of variable domains. *J. Mol. Biol.* **186**, 651–663.
47. Miller, S. (1990). Protein–protein recognition and the association of immunoglobulin constant domains. *J. Mol. Biol.* **216**, 965–973.
48. Chothia, C., Gelfand, I. & Kister, A. (1998). Structural determinants in the sequences of immunoglobulin variable domain. *J. Mol. Biol.* **278**, 457–479.
49. Stoyanov, O., Kister, A., Gelfand, I., Kulikowski, C. & Chothia, C. (2000). Geometric invariant core for the C–L and C–H1 domains of immunoglobulin molecules. *J. Comput. Biol.* **7**, 673–684.
50. Sobolev, V., Sorokine, A., Prilusky, J., Abola, E. E. & Edelman, M. (1999). Automated analysis of interatomic contacts in proteins. *Bioinformatics*, **15**, 327–332.
51. Sobolev, V., Wade, R. C., Vriend, G. & Edelman, M. (1996). Molecular docking using surface complementarity. *Proteins: Struct. Funct. Genet.* **25**, 120–129.
52. Gelfand, I. M. & Kister, A. E. (1995). Analysis of the relation between the sequence and secondary and three-dimensional structures of immunoglobulin molecules. *Proc. Natl Acad. Sci. USA*, **92**, 10884–10888.
53. Stanfield, R. L., Takimoto-Kamimura, M., Rini, J. M., Profy, A. T. & Wilson, I. A. (1993). Major antigen-induced domain rearrangements in an antibody. *Structure*, **1**, 83–93.
54. Connolly, M. L. (1983). Solvent-accessible surface of proteins and nucleic acids. *Science*, **221**, 709–713.
55. Gerstein, M. & Altman, R. B. (1995). Average core structures and variability measures for protein families—application to the immunoglobulins. *J. Mol. Biol.* **251**, 161–175.
56. Gelfand, I. M., Kister, A. E. & Leshchiner, D. (1996). The invariant system of coordinates of antibody molecules: prediction of the "standard" C-alpha framework of V–L and V–H domains. *Proc. Natl Acad. Sci. USA*, **93**, 3675–3678.
57. Gelfand, I., Kister, A., Kulikowski, C. & Stoyanov, O. (1998). Geometric invariant core for the VL and VH domains of immunoglobulin molecules. *Protein Eng.* **11**, 1015–1025.
58. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.
59. Dodge, C., Schneider, R. & Sander, C. (1998). The HSSP database of protein structure-sequence alignments and family profiles. *Nucl. Acids Res.* **26**, 313–315.
60. Mirny, L. & Shakhnovich, E. (2001). Evolutionary conservation of the folding nucleus. *J. Mol. Biol.* **308**, 123–129.
61. Henrick, K. & Thornton, J. M. (1998). PQS: a protein quanternary structure file server. *Trends Biochem. Sci.* **23**, 358–361.
62. Miller, S. (1989). The structure of interfaces between subunits of dimeric and tetrameric proteins. *Protein Eng.* **3**, 77–83.

Edited by J. Thornton

(Received 26 March 2004; received in revised form 23 June 2004; accepted 28 June 2004)