Peptide segments in protein-protein interfaces

ARUMAY PAL[1], PINAK CHAKRABARTI[1,†], RANJIT BAHADUR [1,2], FRANCIS RODIER[3] and JOËL JANIN[2,3]
[1]*Department of Biochemistry, Bose Institute, P-1/12 CIT Scheme VIIM, Calcutta 700 054, India*
[2]*Institut de Biochimie et Biologie Moléculaire et Cellulaire, UMR8619 Bât. 430, Université Paris-Sud, 91405-Orsay, France*
[3]*Laboratoire d'Enzymologie et de Biochimie Structurales, UPR9063 CNRS, 91198-Gif-sur-Yvette, France*

[†]*Corresponding author (Fax, 91-33-2355-3886; Email, pinak@boseinst.ernet.in)*

An important component of functional genomics involves the understanding of protein association. The interfaces resulting from protein-protein interactions – (i) specific, as represented by the homodimeric quaternary structures and the complexes formed by two independently occurring protein components, and (ii) non-specific, as observed in the crystal lattice of monomeric proteins – have been analysed on the basis of the length and the number of peptide segments. In 1000 Å^2 of the interface area, contributed by a polypeptide chain, there would be 3.4 segments in homodimers, 5.6 in complexes and 6.3 in crystal contacts. Concomitantly, the segments are the longest (with 8.7 interface residues) in homodimers. Core segments (likely to contribute more towards binding) are more in number in homodimers (1.7) than in crystal contacts (0.5), and this number can be used as one of the parameters to distinguish between the two types of interfaces. Dominant segments involved in specific interactions, along with their secondary structural features, are enumerated.

[Pal A, Chakrabarti P, Bahadur R , Rodier F and Janin J 2006 Peptide segments in protein-protein interfaces; *J. Biosci.* **32** 101–111]

1. Introduction

Protein-protein interactions pervade all cellular processes of a living organism – namely, in cellular architecture, signal transduction, biosynthetic and degradation pathways and their regulations, immune response etc. – and depend upon precise recognition involving two or more proteins. Indeed, the biological function of a protein can be seen as defined by the context of its interactions in the cell (Eisenberg *et al* 2000). Inappropriate protein-protein binding can lead to disease. Thus the unraveling of the underlying principles governing protein association is central to the construction of protein networks that define cell biology (Edwards *et al* 2002; Aloy *et al* 2004). Various studies have dealt with features that characterize biomolecular recognition (Chothia and Janin 1975; Argos 1988; Lawrence and Colman 1993; Young *et al* 1994; Jones and Thornton 1996; Tsai et al 1997; Lo Conte *et al* 1999; Glaser *et al* 2001; Chakrabarti and Janin 2002; Brinda et al 2002; Bahadur *et al* 2003, 2004; Ofran and Rost 2003; Mintseris and Weng 2003; De *et al* 2005; Saha *et al* 2005). Parameters that can distinguish the contact surface from the rest of the protein surface have been used to identify the interaction sites (Jones and Thornton 1997; Neuvirth *et al* 2004).

Protein antigenic sites (epitopes that are recognized by antibodies) could be generally confined to continuous motifs of about 8–24 amino acid residues, or may occur as short isolated regions along the chain that are brought in close proximity on the surface of the 3-dimensional structure of the protein (Atassi 1984). Thus the protein surface can be mapped in terms of segments along the chain that constitute the overall binding site. Indeed, the presence of a long stretch of contiguous interface residues may suggest the possibility of using the corresponding peptide as a mimic for the binding region of the molecule in its interaction with its biological partner. The synthesized peptide can then be used to prevent the association between the two protein molecules (Gibbs 2000). The fact that the peptides can effectively reproduce the conformation and electronic properties of the functional native protein epitopes has been demonstrated

Keywords. Binding site; interfacial peptides; peptide inhibitor; protein-protein interactions; specific and non-specific interactions

in the case of SH3 (Src Homology 3) domains, which are found in a wide variety of unrelated proteins many of which are involved in signal transduction, suggesting a role in protein-protein interactions (Pawson 1995). These interact with ligand proteins that contain proline-rich sequences that share the PxxP motif (x being any amino acid). Synthetic proline-rich peptides have been found to bind to the domains over a range of up to 7 residues in a polyproline type II helix conformation (Pisabarro and Serrano 1996). Though the binding affinity and specificity of the peptides may not fully represent the *in vivo* situation, the fact that a stretch of peptide can effectively substitute for the whole protein is noteworthy.

The interface can be dissected into core and rim regions (Chakrabarti and Janin 2002; Bahadur *et al* 2003) with the residues constituting the former being more conserved evolutionarily (Guharoy and Chakrabarti 2005) and also contributing more towards the binding of free energy (Bogan and Thorn 1998). Thus the distribution of core residues along the polypeptide chain should also indicate if any particular region in the chain is more important for binding. The segmentation pattern of polypeptide chain containing the interface residues and the location of core residues along the chain are analysed in this article.

Protein-protein interactions can broadly be considered in two groups. While some interactions form stable complexes resulting in permanent, multi-protein structures, others are of a transient nature. Most homodimers belong to the former group, while protein complexes, such as enzyme-inhibitor, antibody-antigen, etc. to the latter (Bahadur *et al* 2004). Apart from these specific associations, protein-protein interactions are also involved in the nucleation and growth of protein crystals. Though non-specific in nature – as seemingly random patches of surfaces are involved in lattice contacts in different polymorphs (Crosio *et al* 1992) – these contacts provide a rich repertoire for understanding the crystallization process and a benchmark for studying the physiological interactions (Janin and Rodier 1995; Carugo and Argos 1997; Dasgupta *et al* 1997). For example, a question relevant to the present-day high-throughput structural proteomics is – if the structure of a given protein has a large contact with a two-fold symmetry related molecule in the crystal lattice, is the physiological state of the molecule dimeric? Different physicochemical parameters can be used, normally in combination, to address this issue (Ponstingl *et al* 2000; Mintseris and Weng 2003; Bahadur *et al* 2004; Rodier *et al* 2005; Saha *et al* 2005). Some of the discriminating factors are the size of the interface and its non-polar component, packing density of the interface atoms and the fraction that is fully buried, amino acid composition, the types of residues involved in self-contact, hydration etc. It would be of interest to see if the features of the peptide fragments that make up the whole interface can also distinguish between the specific and non-specific interfaces.

Table 1. PDB entries for 204 protein complexes.

1A2K	1A2Y	1ACB	1AVA	1AVW	1AXI	1AY7	1AZZ	1B0N	1BJ1
1BLX	1BRS	1BTH	1C1Y	1CA0	1CGI	1CHO	1CLV	1CSE	1CXZ
1CZY	1D4V	1DAN	1DF9	1DFJ	1DHK	1DKD	1DS6	1DTD	1DVF
1DZB	1E44	1E96	1EAY	1EER	1EFN	1EFU	1EMV	1EUV	1EWY
1F34	1FIN	1FLE	1FLT	1FNS	1FS1	1FYH	1GG2	1GL1	1GL4
1GOT	1GUA	1H1R	1H2K	1H2T	1HIA	1I7W	1IAR	1IBR	1ICF
1IGC	1IIL	1IJE	1IOD	1J2J	1J34	1JBU	1JDH	1JDP	1JIW
1JPS	1JTG	1JTH	1JW9	1JYO	1JZD	1K9O	1KB5	1KI1	1KSH
1KTZ	1KXV	1KZ7	1L2I	1L4D	1L6X	1LFD	1LK3	1LPB	1LQV
1M4U	1M9E	1MBX	1MCT	1MCV	1MEL	1MLC	1MZW	1NCA	1NF3
1NL0	1NM1	1NMB	1NU9	1NW9	1O6S	1OC0	1OEB	1OEY	1OFU
1ONQ	1OO0	1OPH	1ORY	1OSP	1OY3	1P5V	1PDK	1PPF	1PXV
1Q1S	1Q40	1QAV	1QTX	1R0R	1R17	1R3J	1REW	1RJ9	1RJC
1RKE	1RP3	1S1Q	1S6C	1SBB	1SBW	1SG1	1SGP	1SHW	1SKO
1SLW	1SQ2	1STF	1SV0	1SVX	1T0F	1T0J	1T0P	1T8O	1TA3
1TAB	1TAW	1TGS	1TH8	1TO2	1TTW	1TX4	1TY4	1U0S	1U6H
1U8T	1UAD	1UJZ	1UKV	1UNL	1US7	1USU	1UXS	1UZX	1V18
1V74	1VF6	1VG0	1VPP	1W98	1WEJ	1WMH	1WMI	1WQJ	1WWW
1XB2	1XD3	1XG2	1XL3	1XT9	1XX9	1YCS	1YDR	1YQV	1YRO
1Z7K	2BF8	2BO9	2KAI	2NGR	2PCC	2PRG	2PTC	2SIC	2TEC
2TRC	3FAP	3TPI	4HTC						

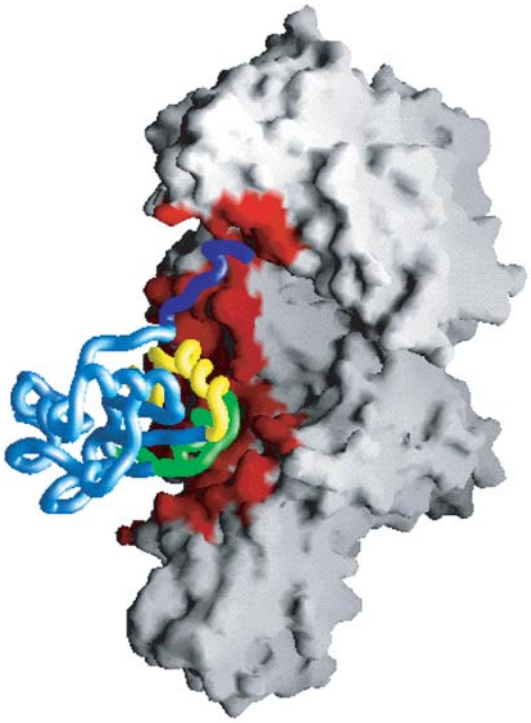

Figure 1. Three segments (residues 5–14 in blue, 50–69 in green and 90–110 in yellow; the non-interface region being in cyan) shown against the backbone tube of small CBP20 bound to large CBP80 (whose surface is shown in grey and the interface is coloured red) in the heterodimeric human nuclear cap-binding complex (CBC) in the PDB file 1H2T (Mazza *et al* 2002).

2. Datasets and methodology

The datasets of 122 homodimers and 188 large interfaces (with size > 800 Å^2) formed in the crystal lattices of monomeric proteins have already been described (Bahadur *et al* 2003, 2004). We updated the list of protein-protein complexes used in an earlier study (Chakrabarti and Janin 2002) and the present non-redundant set (table 1), compiled from the Protein Data Bank (PDB) (Berman *et al* 2000), contains 204 complex structures that have been determined to a resolution of 2.5 Å or better. Interface residues were identified based on the loss of the solvent accessible surface area in the complex relative to the value in the individual components. From symmetry consideration, only one subunit of homodimers and 103 crystal contacts that are related by a 2-fold symmetry (crystallographic or non-crystallographic) (Bahadur *et al* 2004) were used in the analysis. For the rest, each subunit was taken up individually, along with its share of interface residues and the interface area.

We define an interface segment as a stretch of residues that starts and ends with interface residues and may contain intervening non-interface residues, but only in stretches of not more than four. Jones and Thornton (1996) used a slightly different condition (> 5 intervening non-interface residues to separate interface residues into segments). While considering the length of the segment, only interface residues in it are counted. The secondary structures of proteins were determined using DSSP (Kabsch and Sander 1983) and the surface representation of the molecules was created using GRASP (Nicholls *et al* 1991).

3. Results and discussion

The number of polypeptide chains that has been used is 122 in homodimers, 408 (2 x 204) in complexes and 273 in crystal contacts (103 for 2-fold symmetry contacts and 2 x 85 = 170 for non 2-fold symmetry contacts). A typical example of peptide segments in the interface is shown in figure 1.

3.1 *Length and the number of segments*

The number of segments can vary from 1 to as high as 20 (in case of homodimers) (figure 2). The average numbers are quite similar in all the three categories of interfaces and close to 5.2 (±2.6) reported earlier for a dataset of 32 homodimers. However, as there are large differences in the size and the number of residues, especially between homodimers and crystal contacts, comparison should be made only after the numbers are normalized relative to these factors. Thus, per 1000 Å^2 of the interface area contributed by a subunit, the homodimers have on average 3.4 (±1.6) segments, almost half the value of 6.3 (±2.2) for the crystal contacts, the complexes have an intermediate value closer to the latter (figure 3a). Only 12% of the homodimeric interfaces have more than 5 segments per 1000 Å^2, whereas for crystal contacts the number is as large as 72%. Similar trends can also be seen when the normalization is done

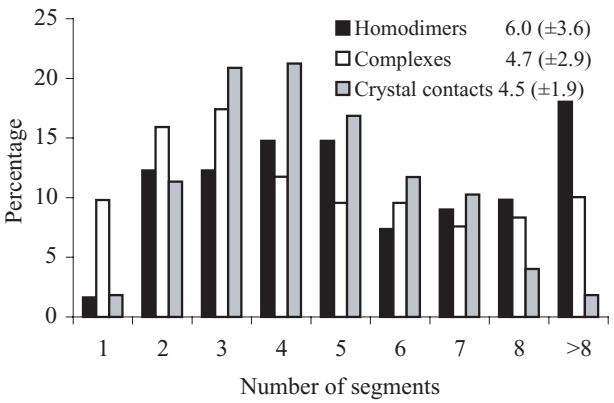

Figure 2. The distribution of the number of segments in interfaces of different kinds.

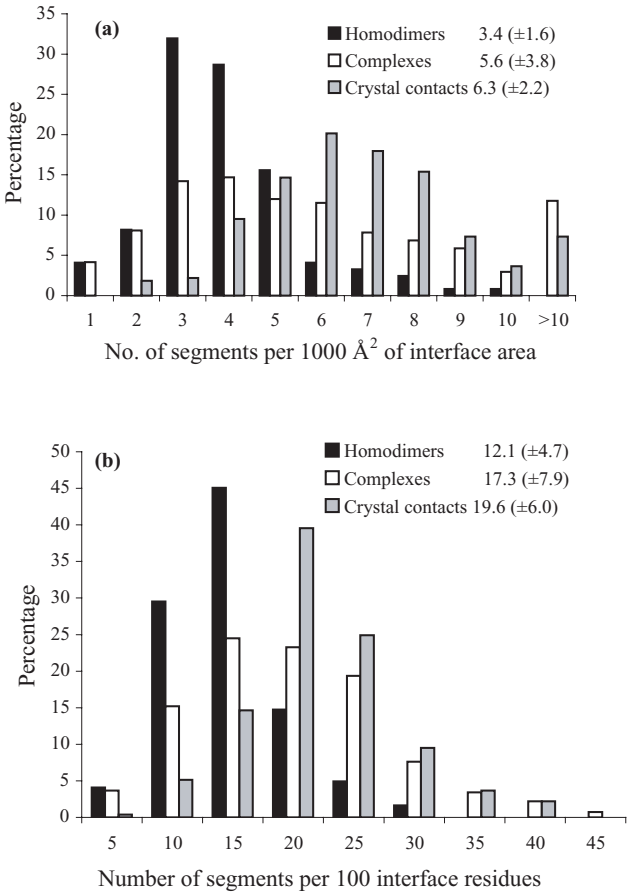

Figure 3. The distribution of the number of segments in each interface **(a)** per 1000 Å^2 of interface area, and **(b)** per 100 interface residues.

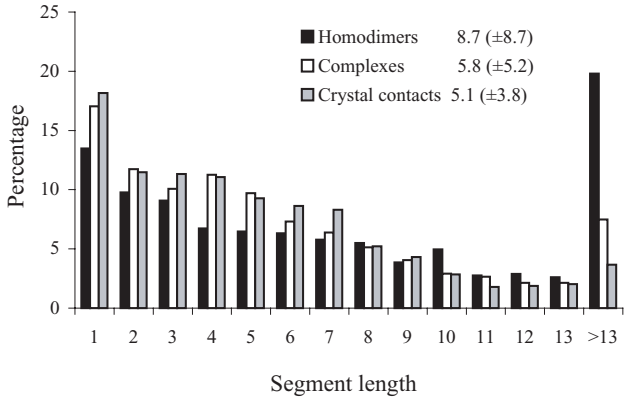

Figure 4. The distribution of the number of residues per segment.

relative to the number of interface residues (figure 3b). (It may be mentioned that the scale factors of 1000 and 100 used in figure 3 are arbitrary – the average contribution of an interface residue to the interface area is ~40 Å^2). The reason why the normalized number of segments is the smallest for the homodimeric interfaces is because these also have the longer segments (figure 4) – 33% of these have length > 9; a comparable length is possessed by only 12% of interfaces formed by crystal contacts. Crystal contacts are usually more fragmented (and thus more discontinuous) than the homodimeric contacts (Bahadur et al 2004).

3.2 *Features of the core segments*

Core residues have been defined as the ones which contain one or more fully-buried interface atoms (Chakrabarti and Janin 2002; Bahadur *et al* 2003) and these usually contribute more towards the free energy of binding (Guharoy and Chakrabarti 2005). As the name implies these residues form the core of the biological interfaces, being surrounded by a rim of residues with lesser energetic contribution towards binding. To analyse the distribution of core residues among the segments we designate those that have at least 70% of the interface residues belonging to the core as core segments. On average the biological interfaces have 1–2 core segments, the corresponding number for the crystal interfaces being 0.5 (figure 5); of all the peptide fragments in the latter only 35% are core fragments.

Assuming that the core segments contribute more towards binding one can study if the percentage composition of interface residues in these segments is any different from that in the non-core segments (figure 6). Euclidean distances between the compositions indicate that the difference is the maximum in crystal contacts. Indeed, for the crystal interfaces the contribution of the charged residues, such as Lys, Asp and Glu is lesser in the core segments as compared to the non-core segments. If we look into the top panel comprising of core segments, the contribution of Leu and

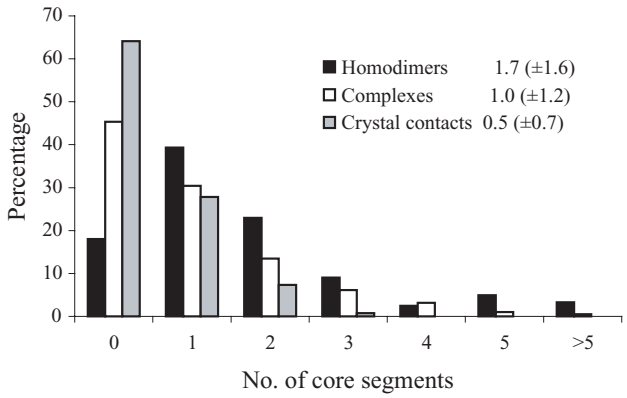

Figure 5. The distribution of the number of core segments. A core segment has at least 70% of its residues that are part of the interface core (residues that contain fully buried interface atoms).

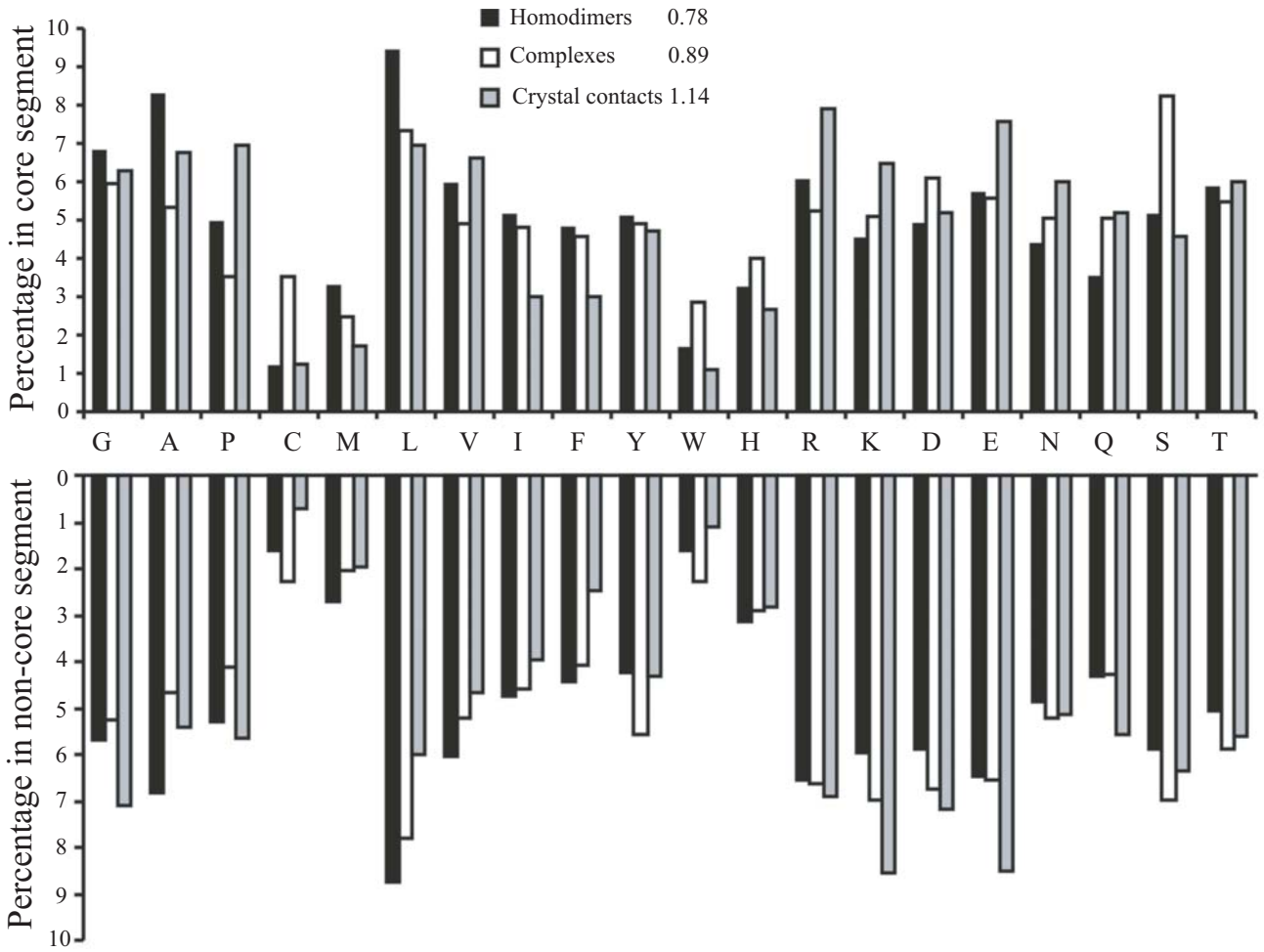

Figure 6. The percentage composition of interface residues present in the core (top panel) and non-core (bottom panel) segments. The Euclidean distance, Δf between the values in the top and bottom panels is provided against the name of each dataset, and is given by

$$(\Delta f)^2 = 1/19 \sum_{i=1 \text{ to } 20} (f_i - f'_i)^2,$$

where f_i and f_i' are the percentage contribution of amino acid type i to the interface belonging to the core and the non-core segments, respectively.

Ala is the maximum for homodimers, and this could be due to the preponderance of such residues in what has been termed as self-contacts (Saha *et al* 2005) involving residues lining the 2-fold axis. Crystal contacts have more of charged residues, Arg, Lys and Glu, in particular.

3.3 *Distinguishing between homodimers and crystal contacts*

Assessing the oligomeric state of a protein from its X-ray structure is a non-trivial problem and we examined the usefulness of the segmentation features of the interfaces as a tool to achieve this. The two parameters developed here, viz., the number of segments (Ns) in 1000 Å^2 of the interface area and the number of core segments (Ncs) are plotted in figure 7, and we looked for boundaries that can separate cases arising out of homodimeric and crystal interfaces. Interestingly enough, if one counts the total number of points lying in the two regions (Ncs $\geq$ 3) and (Ncs < 3 and Ns < 5), the homodimer dataset yields 89%, while for crystal contacts it is 29%. Thus, while the two parameters together can identify quite a high percentage of homodimers their performance is not very good in excluding the crystal contacts. If the second region is reduced to (Ncs in the range 1–2 and Ns < 5) the discrimination against the crystal contacts increases (only 11% of the points are contained in the two regions), but the percentage of homodimers is also reduced to 74%. Thus the first set of conditions to identify the homodimers gave a success rate of 89%, though it also

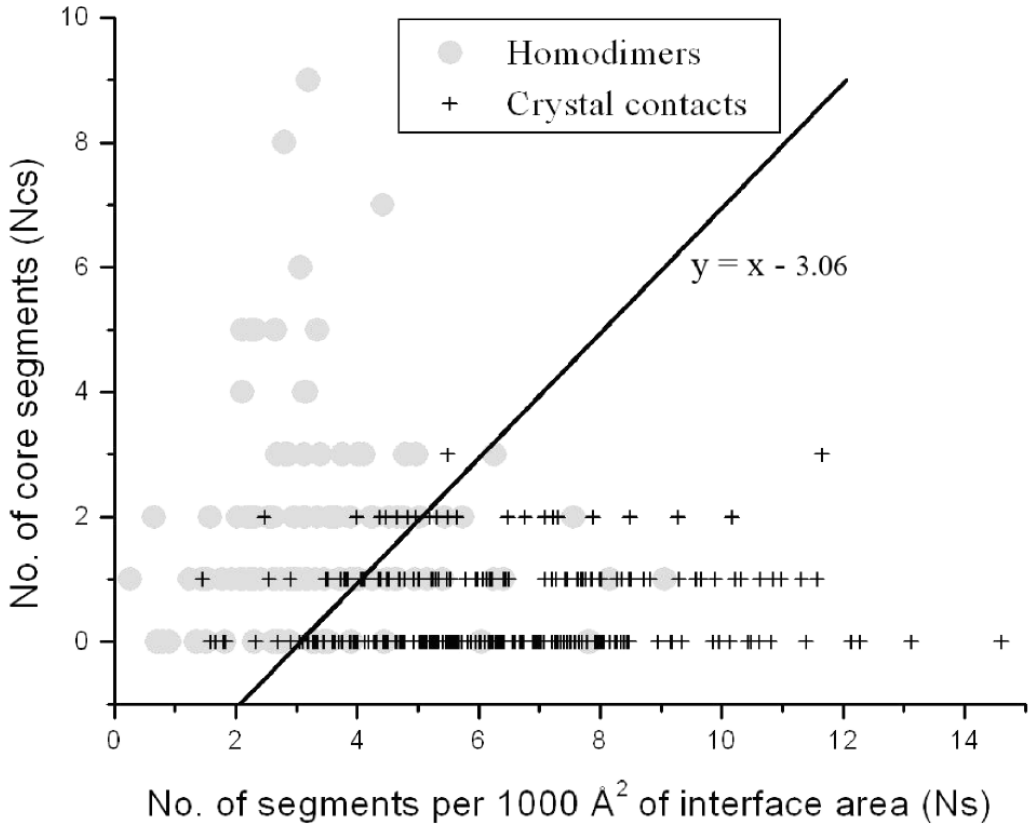

Figure 7. Joint distribution of the number of segments in 1000 Å^2 of the interface area and the number of core segments present in the interfaces of homodimers and crystal contacts. A line has been drawn to demarcate homodimers from the crystal contacts and its equation is shown.

picked up a large number of false positives; with the second set, the false positives were less, but the success rate also came down to 74%.

Instead of defining rectangular regions, one can also use a linear classification scheme. 80% of the homodimeric interfaces lie above the line shown in figure 7 and 91% of the crystal interfaces lie below it. Thus the segmentation features can be a useful tool to distinguish between homodimers and crystal contacts, though the success rate is lesser than ~94% obtained when a combination of three parameters is used for discrimination (Bahadur *et al* 2004).

3.4 *Dominant segments*

We went beyond the core segment to identify the dominant segments, defined as the ones, which contain at least 70% of the total core residues in the whole interface. Table 2 provides the sequence and secondary structural information on such segments. The homodimers with 209 core segments have only 28 dominant segments, whereas the corresponding numbers for the complexes are 393 and 116 (of which a disproportionately large number, 76, is contributed by the second component, which in many cases is the inhibitor in an enzyme-inhibitor complex), respectively. In both homodimers and complexes, 76% of the dominant segments come from interfaces that are made up of 1 or 2 segments in figure 2. Figure 8 shows two extreme examples where virtually the whole polypeptide chain forms the interface. The dimer in figure 8a is formed by the intercalation of two polypeptide chains; in figure 8b, 61 residues (7–67) of RelB are observed in the crystal structure and barring a few residues in the beginning, 57 residues constitute the segment (with 44 residues in the interface), which wraps around the molecular surface of RelE.

4. Concluding remarks

In this paper we report a study of the segmentation pattern of the polypeptide chain in the interface region. When viewed in the backdrop of non-specific interactions observed in the crystal lattices of monomeric proteins, such features provide important lessons for understanding the specificity

Table 2. List of dominant segments in biological interfaces (PDB file name, the chain ID, the residue number of the first residue in the segment and the sequence are provided).

(a) Homodimers (28 segments)

1A3C A 112 **RtVRagMDalVDv**GRP*S*SIQLavlvDrghR*ELP*IRa*D*
1A4I A 173 Rs*Kivg***APmHDllLW***nN*ATVTTCH*sKTAhlde***Evn***K***
1AA7 A 75 QRR**RfvQNa***LN***gnG**dPNnMDkaVKlYRkl***K***
1B3A A 2 PYSSDTTPSS*F*AYI
1B67 A 3 E*LP***iaPIgRIIkn***AGAERVS***dDAriALakVLEeMGEeIASeAV***
1BAM A 116 N**ISSaHRsMNKlLLglk***H***
1BIF A 218 *MDvgQ*SYVVNRVA*D***HiqSRiVYyl***MN***
1BSR A 4 **AAakFERQH**MDSG*NsPSS***SsnYcNLMmSS***RK*MtQgks*Kpv*NTFVH*ES***LadVK***
1BXG A 2 **IDsa***LN*WD*GE*MTVTRFd*Sktga*HfvirLds*T***
1CDC A 4 GTVWG*ALG*HGINLN*IP*NFqM*tD*DIDEVRWE*RG*sTLVAEF*KrKM*kPFLK*sga*Fe*I LANG*DLK*IK*NL
TRDDSGTYNVTV*YST*NGTrILdKALDLRILE
1CG2 A 230 *AGA***PElgv**N**aLVeaSDlVLrtMni**DdKaknl*RfNWTIAKAG*NVSNII*P*asat*LNADvR
1DXG A 12 *C*GQVVKVLEEG*GG*TLVC*Cged*M
1E98 A 61 *S*dVE**dHSvhLLfSAnrW***Eq***VP***
1FIP A 26 P**LRdSVKQALKNYF***aQLngq*DV*ND***LYELVLaEVEQALlDMVmqY***
1HJR A 68 VFmAKN*ADs***ALkLGQaRGVaIVaaV***NqelpVfE***
1JHG A 14 **QRhqEWLRFVDLLKNAYq***NdL***HLPLLNLM***LT***PdEREALGTRVRiIE***
1KBA A 46 SpQFr*Sn*YRSLLSS*TT*
1PGT A 59 *D*LTLYQ**sNTilRHlgRTlgl***Y*gkd**QQeAALvDMVnDGveDlrC***
1REG A 85 EIV*PG*QRTFMK
1RPO A 2 T**kQEkTAlnMArFIrsQTlTLleKLnE***L***aDAADEQadICeSLhdHAdELyrSClaRF***
1TC1 A 51 *LKgs***FMfTADlCRAlCD***fn*VPVRMEFiCvS***
1UTG A 21 **YetsLk***EF*EPD**dTMkdAGmQMkKVL***dSLP***qtTreNImKLTeKIV***kS***pLC***
2ARC A 136 **QiinAgqg**E*GrY***SeLlaiNLlEQllLRrME***
2HDH A 195 HPVSCK*dTpgFI***VnRLLvPYlxEAirLye***RGDA*ske**DIdTAxkLGA**GY*Pxg***PfeL***
2OHX A 283 *QEAY*gvsvIV*G*VPP*dS*QNLSMN*P***MlLL***SG*RTWKGA*I***
2SPC A 40 **LikKHedFDkAIngHEqkIaaLQtvAdqLIaQNHYASnLVDeKRkqVLeRWrHLKeGLIeKRsR***LG***
3SDH A 64 **DKlRGHsITlmYalQNfiD***QldN***pDDlvCVveKFaVNhit***R***
3SSI A 6 LYaPsALVlTVGkgvs*Att*AAPERAVTLT

(b) Complexes (116 segments)

1A2Y C 116 *KGTD***VQaw***IRGs*RL
1ACB I 42 PVTLDLR*Yn*RvRvF
1AY7 B 29 *YYGeN***LDAlWDClTG***W***
1B0N A 74 LD**sEWEKLVRDAMT***Sg*VS**KkQFReFLdyQKwrK**
B 9 FELD**QEWVELMVEAKE***AN*IS**PeEIRkYLll**N
1BJ1 W 79 QIMRIK*PHQ*GQHIGEM
1BRS D 27 Pe*Y*Yge*N***LDAlWDAlTg***W*vE
1C1Y A 21 **VqfVQ**gIfVe*K*yDPt*IE*DSYRK
1CGI I 10 Y*NEL**NGS*TYEYRP
1CSE I 35 YfLp*Eg*SPVTLDLR*yn*RvR
1CZY D 204 PQQA*T*DD
1DKD E 602 WMTT*pW*GFLHP
1DS6 B 22 GNYkPPPQKs**lkeLqe***MDkd*D**ESLiKYKkTL***L***
1DVF D 95 Kv*IYYQGR*
1E96 A 21 **ISyTT**NAFPG*E*yIpt*V*fdnYS
1EAY A 90 Ak**KenIIaAAQ***ag*ASGY
C 202 D*I*a**eDDitAVl***CFVIE***

```
1EFN B  71 RPqVPlRpmtyKaavDLshFlkE
1EMV A  22 ICNAdTSSEeeLVklVThfEemteHPSGSDlIYYP
     B  70 NLNPSNkSSvsKgYSpFTPKNqQvgGrKVY
1EUV B  61 DSlRfLYDGIRIQAdQtpedLDmednDiiEaHREQIGG
1FLE I  18 IILIRSAMLNpPnR
1FNS A 627 PQRmSRNFVR
1FS1 A 109 WdSLPdeLllGIFsCLClpElLKVsGVckrW
1GG2 A 197 KmFdVgaQrSErKKWiHCFEgV
1GL1 I   9 TfkdKsNTsRSGAdgkSAASTLKASPN
1GOT A 193 RsFdVggQRSErKKWiHCFEgV
1GUA A  21 VqfVQgIfVdEyDptIEDSYR
1HIA I  25 NEVHSRIRsKygL
1JBU X   1 EEWEVLSWTWEtSER
1JDH B  12 LGANDELISFkdEGEqEEksSENSSAErDLaDVKSSLV
1JTH A  19 DqlADeSLesTRrMLqlVEeSKdaGIrTLvmLDeQGeqLErIEegMDqINkdMK
     B 194 EieTRhsEIiKLenSIrELhdMfmDMamLVesQgeMIdRIeyNVeHAvdYVerA
1JDP H   5 GSFGLKLDRIGSMSGLGS
1K9O I 342 AAAAnAFGIVPKSLIL
1KSH B  90 WfFEFgFVIPNSTNTWQSLiE
1KTZ B  49 SITSISE
1L2I A 355 VhmINwaKrvpgFvdltLHdQVhlLE
     C 687 HKILHrLLQD
1LFD A  27 NGNMYKSIL
     B 221 IqlIQnHfVdKyDPTIEDSYRK
1LQV B  74 LQfhGLvRLvhQ
     D   2 NSFLssL
1M4U A  27 MqHYLHIRPA
     L  73 FPLNSymnAtNhaIVqTL
1M9E C  85 PVAAGPIAPGQ
1MEL A 101 TIYASYYEbgH
1MCT I   1 RISPRIWM
     I   1 PSTLEYMR
1MZW B 108 KaslRalGEPITLFGEgpaeR
1NF3 C 131 IVISMPQDFRPVSSIIdVDilpEthR
1NW9 A 292 LgegdKVKcfhCGGGLTDWKpsEdpweQhaKWYPGCKylLeqkgQEyINNiHLTH
1NL0 G   4 GKLxxFVQGnlxR
1OC0 B   9 CTEGFnVdkksQsDELssYYQS
1OEB D   2 APSIDRsTKPPL
1OEY A 353 TlKvHYKYTVVsK
1OPH A 350 AmfLEAIPRSIP
1ORY B2479 NVDFAKeMTEFTKyQIRmQSgVAMLAQANALPQLVLQLLR
1Q1S B 119 DAQHAAPPKKKRKVE
1QTX B   1 RRKWQKTGHAVRAIGRLSsS
1REW B  48 PFPLADHLNSTNhaIVqTLvNSVNS
1R0R I  10 EYpKPACTLEYR
1R3J C  45 YLaVLAERGAPGAQlITYPR
1RP3 B  35 DKvtLSkIAQeLSknDvEekdLEkKVkeLKekIEkgeYEVSDEKVVKGLIEFFT
1S6C B 217 MAAGVAAWLPFarAAAiGW
1SBB A  47 HysYGAGSTEKgdiPdgYKASrPSQ
1SBW I  15 SCRcTKSIPPqchC
1SHW A 118 FSeKfQLFTPFSLgfEfR
1SGP I  10 EypKPAdTAEYR
1SKO A  50 FlsTFaLAtdQGsKLGLSKNKSIICYYNTYqvvQ
```

1SLW A 78 SSPVSTMMHSP
1SQ2 N 85 LGV*AGg*YcD**yALcs***SRY*AE
1SV0 D 133 **QK**fP*M*N**gKAlcL**Ms**ldMylCR**
1T0F C 505 IKVVk**PsdWds***LPdTD***LRyiYs***qRqpeKT***MHerLKg***kGV*IVD**MASL***FK*
1T0J C 429 **QLEEDLkGYLDWIt**Q
1TA3 A 225 *DKSHq*WvH**PKNvYY***G***VapV**
1TAB I 21 QCSCT*KSMP*PKcRcS
1TAW B 11 TG*Ph*RAMIS
1TGS I 10 TS*EVSg*CPKIYNP
1TO2 I 49 IVLP*vGT*IVTKEYR*id*RvRlF
1TTW B 33 VSTQAIT*sd***ERrFAYaV***L*
1TY4 A 120 **GtIFekKHaeNFeTFceQLl***aV***Pri**sf**SPyqDVvRTVgn**AQT*dQcpsS***YGRlIGLisF**
C 48 Ss**IGYEIGSKLAaxCDDFDaQxxSY***Sa*H
1U0S A 186 K*eGTQL***kSARiYLvfHkleE**
Y 82 S*AMGQ***QamVIeaIK***ag*AKDFIV*KPf*Q**psRvvEAlnK**
1U6H B 853 SR**KLLsAAKILAdATAKMVEA***AKG*
1U8T A 90 AK**KenIIaAAq***ag*ASGWVV
E 4 SILs**QaeIDaLL**N
1UAD A 44 Ept*K*ADSYRK
C 11 TGISPNEG*ipw*TkvTiRGE
1UJZ B 513 *FS***RNNnDRmk***Vg*KapQTR**TQd**Vs*gK*rRSF
1US7 A 99 **SgtKAfmEAlS***AGAD***vsMiGQF***G*
1USU A 433 *TQn***RA**
1VF6 A 12 V**lqVLdRLkmKLqeK***gDt**SQneKLsmFyeTLk**SpLFnqILtlQQsiK**
C 126 **DLfsSLkHIq***hTLVD***sQSqeDIsLLlqLVq**n**rDFqNAFkIHNaV**T
1VPP X 1 RGWVEICAA*DDYGrC*
1WMH A 60 KWi*DEEg*DPCTvS*S***qLEleEafRLyeL**
B 15 IvEVKskf*dAEFRRFAL*
1WMI A 1 *MTYRVK***IHKQvVk***aLqs***lpKahYRrFLefRD***ilEYEP***VpRe**kfdVIKLE*GT*GDLDLYRaRlg
*dyRvIySvNW**kdKV*iKILKlKPr*GRA*
B 11 **KelERlKVEiQRlEAM**l*MPEERDEDIt***eeEIaELLeLA***RDEd**PEN**WIDAEe*LPEPED*
1WQJ I 2 PETLc**gaElvDAlQFVcGDR**GFY*F*
1WWW W 2 SSH*P***IFhR***GEFSvs*DSVSVW
1XG2 B 66 **KlkGRyETcsENyaDaiDSlgQakqFlts***gD***yNSlNIYaSAaFDGaGTceDS**fE
1YCS A 239 N*sSCMggmNR*RP
1YDR I 6 **TYaDFI***aSgRTg*RRNAIHD
1Z7K B 20 MILCNKALNP
2BF8 A 3 **NiavQriKRefK**
2KAI I 11 Tg*PC*KARII
2PCC A 31 **RE***dDE***ydNY**
B 5 **KgaTLfKTRc***lQC*
2PRG C 685 E*RH***kILHrLLQ***eGS*PsDI
2PTC I 11 Tg*PC*KARIIR
2SIC I 65 Rg*ED*VMcPMVY
2TEC I 35 YfLp*EG*SPVTLDLR*Yn*RvRvF
3FAP B 127 **YF***ge*R
3TPI I 11 Tg*Pc*KARIIR

The non-interface residues in the segment are in small letters; the residues in helices [of any type, with DSSP designation (H or G)] are in bold; those in β-strands (E or B) are underlined and the ones in turns (S or T) are in italics. An 'X' or 'x' in the sequence indicates a modified residue (such as selenomethionine, γ-carboxyglutamic acid, etc.).

of macromolecular recognition. Earlier we showed how the interface size and the fraction of fully buried atoms could be used to discriminate the biological interfaces from the non-biological ones (Bahadur *et al* 2004). Here the number of segments has been used towards the same goal, though the interface size and the information on the fully buried

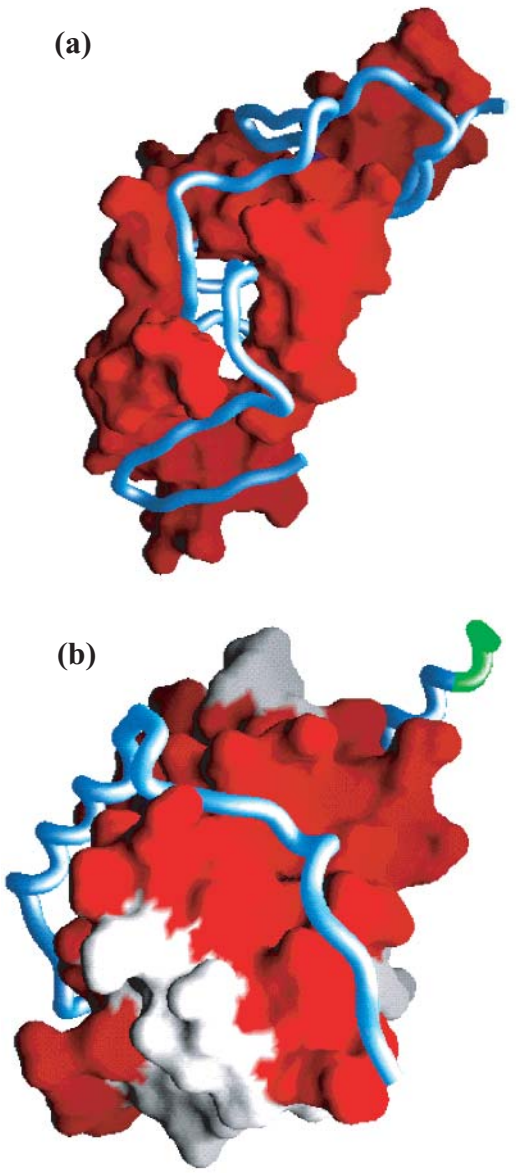

Figure 8. Examples of dominant segments (imaged as backbone tube against the surface representation of the other subunit): **(a)** homodimer of the N-terminal domain of the lymphocyte cell adhesion molecule CD2 (PDB code: 1CDC) (Murray *et al* 1995); **(b)** toxin-antitoxin RelE-RelB complex (the latter is wound around the former) (PDB code: 1WMI) (Takagi *et al* 2005).

atoms have also been incorporated in an indirect manner in the definition of the two parameters in figure 7. The control of protein-protein interactions is believed to be the next frontier in pharmaceuticals (Toogood 2002; Arkin and Wells 2004) for designing peptides aimed at inhibiting protein interactions (Fasan *et al* 2004). The dominant segments delineated here could be used as template structures for the design.

Acknowledgements

This work was supported by a grant from the Council of Scientific and Industrial Research, New Delhi. PC is grateful to Université Paris-Sud for a visiting position.

References

Aloy P, Böttcher B, Ceulemans H, Leutwein C, Mellwig C, Fischer S, Gavin A, Bork P, Superti-Furga G, Serrano L and Russell R B 2004 Structure-based assembly of protein complexes in yeast; *Science* **303** 2026–2029

Argos P 1988 An investigation of protein subunits and domain interfaces; *Protein Eng*. **2** 101–113

Arkin M R and Wells J A 2004 Small-molecule inhibitors of protein-protein interactions: progressing towards the dream; *Nat*. *Ret*. *Drug Discov*. **3** 301–317

Atassi M Z 1984 Antigenic structures of proteins: their determination has revealed important aspects of immune recognition and generated strategies for synthetic mimicking of protein binding sites; *Eur. J. Biochem.* **145** 1–20

Bahadur R P, Chakrabarti P, Rodier F and Janin J 2003 Dissecting subunit interfaces in homodimeric proteins; *Proteins* **53** 708–719

Bahadur R P, Chakrabarti P, Rodier F and Janin J 2004 A dissection of specific and non-specific protein-protein interfaces; *J. Mol. Biol***. 336** 943–955

Berman H M, Westbrook J, Feng Z, Gilliland G, Bhat T N, Weissig H, Shindyalov I N and Bourne P E 2000 The Protein Data Bank; *Nucleic Acids Res*. **28** 235–242

Bogan A A and Thorn K S 1998 Anatomy of hot spots in protein interfaces; *J. Mol*. *Biol*. **280** 1–9

Brinda K V, Kannan N and Vishveshwara S 2002 Analysis of homodimeric protein interfaces by graph-spectral methods; *Protein Eng*. **15** 265–277

Carugo O and Argos P 1997 Protein-protein crystal-packing contacts; *Protein Sci*. **6** 2261–2263

Chakrabarti P and Janin J 2002 Dissecting protein-protein recognition sites; *Proteins* **47** 334–343

Chothia C and Janin J 1975 Principles of protein-protein recognition; *Nature (London)* **256** 705–708

Crosio M P, Janin J and Jullien M 1992 Crystal packing in six crystal forms of pancreatic ribonuclease; *J. Mol*. *Biol*. **228** 243–251

Dasgupta S, Iyer G H, Bryant S H, Lawrence C E and Bell J A 1997 Extent and nature of contacts between protein molecules in crystal lattices and between subunits of protein oligomers; *Proteins* **28** 494–514

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

Edwards A M, Kus B, Jansen R, Greenbaum D, Greenblatt J and Gerstein M 2002 Bridging structural biology and genomics:

assessing protein interaction data with known complexes; *Trends Genetics* **18** 529–536

Eisenberg D, Marcotte E M, Xenarios I and Yeates T O 2000 Protein function in the post-genomic era; *Nature (London)* **405** 823–826

Fasan R, Dias R L A, Moehle K, Zerbe O, Vrijbloed J W, Obrecht D and Robinson J A 2004 Using a β-hairpin to mimic an α-helix: cyclic peptidomimetic inhibitors of the p53-HDM2 protein-protein interaction; *Angew. Chem. Int. Ed.* **43** 2109–2112

Gibbs J B 2000 Mechanism-based target identification and drug discovery in cancer research; *Science* **287** 1969–1973

Glaser F, Steinberg D M, Vakser I A and Ben-Tal N 2001 Residue frequencies and pairing preferences at protein-protein interfaces; *Proteins* **43** 89–102

Guharoy M and Chakrabarti P 2005 Conservation and relative importance of residues across protein-protein interfaces; *Proc. Natl. Acad. Sci. USA* **102** 15447–15452

Janin J and Rodier F 1995 Protein-protein interaction at crystal contacts; *Proteins* **23** 580–587

Jones S and Thornton J M 1996 Principles of protein-protein interactions; *Proc. Natl. Acad. Sci. USA* **93** 13–20

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

Kabsch W and Sander C 1983 Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features; *Biopolymers* **22** 2577–2637

Lawrence M C and Colman P M 1993 Shape complementarity at protein/protein interfaces; *J. Mol. Biol.* **234** 946–950

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

Mazza C, Segreff A, Mattaj I W and Cusack S 2002 Large-scale induced fit recognition of an m^7GpppG cap analogue by the human nuclear cap-binding complex; *EMBO J.* **21** 5548–5557

Mintseris J and Weng Z 2003 Atomic contact vectors in protein-protein recognition; *Proteins* **53** 629–639

Murray A J, Lewis S J, Barclay A N and Brady R L 1995 One sequence, two folds: a metastable structure of CD2; *Proc. Natl. Acad. Sci. USA* **92** 7337–7341

Neuvirth H, Raz R and Schreiber G 2004 ProMate: a structure based prediction program to identify the location of protein-protein binding sites; *J. Mol. Biol.* **338** 181–199

Nicholls A, Sharp K and Honig B 1991 Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons; *Proteins* **11** 281–296

Ofran Y and Rost B 2003 Analysing six types of protein-protein interfaces; *J. Mol. Biol.* **325** 377–387

Pawson T 1995 Protein modules and signalling networks; *Nature (London)* **373** 573–580

Pisabarro M T and Serrano L 1996 Rational design of specific high-affinity peptide ligands for the Abl-SH3 domain; *Biochemistry* **35** 10634–10640

Ponstingl H, Henrick K and Thornton J M 2000 Discriminating between homodimeric and monomeric proteins in the crystalline state; *Proteins* **41** 47–57

Rodier F, Bahadur R P, Chakrabarti P and Janin J 2005 Hydration of protein-protein interfaces; *Proteins* **60** 30–45

Saha R P, Bahadur R P and Chakrabarti P 2005 Inter-residue contacts in proteins and protein-protein interfaces and their use in characterizing the homodimeric interface *J. Proteome Res.* **4** 1600–1609

Takagi H, Kakuta Y, Okada T, Yao M, Tanaka I and Kimura M 2005 Crystal structure of archaeal toxin-antitoxin RelE-RelB complex with implications for toxin activity and antitoxin effects; *Nat. Struct. Mol. Biol.* **12** 327–331

Toogood P L 2002 Inhibition of protein-protein association by small molecules: approaches and progress; *J. Med. Chem.* **45** 1543–1558

Tsai C J, Lin S L, Wolfson H J and Nussinov R 1997 Study of protein-protein interfaces: a statistical analysis of the hydrophobic effect; *Protein Sci.* **6** 53–64

Young L, Jernigan R L and Covell D G 1994 A role for surface hydrophobicity in protein-protein recognition; *Protein Sci.* **3** 717–729

*e*Publication: 6 September 2006