

1 Journal of Bioinformatics and Computational Biology
2 Vol. 3, No. 5 (2005) 1–14
3 © Imperial College Press



5 **A QUANTITATIVE ANALYSIS OF INTERFACIAL AMINO ACID
6 CONSERVATION IN PROTEIN-PROTEIN HETERO COMPLEXES**

7 BOOJALA V. B. REDDY* and YIANNIS N. KAZNESSIS*,†

8 **Digital Technology Center and*
9 **†Department of Chemical Engineering and Materials Science*
10 *University of Minnesota*
11 *Minneapolis MN 55455*

12 Received 1 December 2004

13 Revised 18 March 2005

Accepted 28 March 2005

15 A long-standing question in molecular biology is whether interfaces of protein-protein
16 complexes are more conserved than the rest of the protein surfaces. Although it has
17 been reported that conservation can be used as an indicator for predicting interaction
18 sites on proteins, there are recent reports stating that the interface regions are only
19 slightly more conserved than the rest of the protein surfaces, with conservation signals
20 not being statistically significant enough for predicting protein-protein binding sites.
21 In order to properly address these controversial reports we have studied a set of 28
22 well resolved hetero complex structures of proteins that consists of transient and non-
23 transient complexes. The surface positions were classified into four conservation classes
24 and the conservation index of the surface positions was quantitatively analyzed. The
25 results indicate that the surface density of highly conserved positions is significantly
26 higher in the protein-protein interface regions compared with the other regions of the
27 protein surface. However, the average conservation index of the patches in the interface
28 region is not significantly higher compared with other surface regions of the protein
29 structures. This finding demonstrates that the number of conserved residue positions is
30 a more appropriate indicator for predicting protein-protein binding sites than the average
31 conservation index in the interacting region. We have further validated our findings on
32 a set of 59 benchmark complex structures. Furthermore, an analysis of 19 complexes of
33 antigen-antibody interactions shows that there is no conservation of amino acid positions
34 in the interacting regions of these complexes, as expected, with the variable region of the
35 immunoglobulins interacting mostly with the antigens. Interestingly, antigen interacting
36 regions also have a higher number of non-conserved residue positions in the interacting
37 region than the rest of the protein surface.

38 *Keywords:* Protein complexes; docking; molecular recognition; protein-protein interac-
39 tion; sequence conservation; protein evolution.

40 **1. Introduction**

Protein-protein recognition at the molecular level is the basis for numerous phys-
iological processes in the cell. Determining the exact mode of interaction between

2 B. V. B. Reddy & Y. N. Kaznessis

1 two protein molecules is important for gaining the molecular level understanding
2 of these physiological processes. Study of evolutionary conservation and variation
3 of the amino acids in the exponentially growing datasets of protein sequences and
4 their structures provides very useful information about the basis of protein-protein
5 interactions and their importance in life processes.¹

6 Protein-protein interaction sites have specific chemical and physical characteris-
7 tics, all of which contribute to the molecular recognition process.² They are diverse,
8 hydrophobic, planar, globular³⁻⁶ and typically involve large accessible sites where
9 the solvation potential, interface propensities and protrusion of residues cannot be
10 easily distinguished from the rest of the protein's surface.⁷ For transient protein-
11 protein interactions, binding surfaces are found to lack charged groups and have
12 an excess of hydrophobic residues, leading to an amino acid composition that is
13 intermediate to that of the protein interior and surface.⁸ The physical and chemical
14 aspects of subunit interfaces in oligomers have been extensively studied⁹ in order
15 to develop methods for prediction of putative interfaces using protomer structures
16 alone.^{1,7,10,11}

17 It has been long observed that the three-dimensional structural constraints
18 and functional selection of proteins in nature leads to the retention of significant
19 sequence homology between proteins of similar fold and function.^{12,18} In the case
20 of permanent protein complexes of homodimers it was observed that there is signifi-
21 cant residue conservation at the interfaces compared to other parts of the protein
22 surface.⁸ The conservation of amino acids at the interaction site depends not only
23 on the geometry and electrostatic complementarities of the interacting surfaces but
24 also on the context of its functional contribution. The efficiency of the binding
25 of some active sites is allosterically regulated and the site can adapt to muta-
26 tions through extensive structural rearrangements.¹⁹ Thus, the effective functional
27 site not only includes part of the ligand contact site, but also extends beyond it
28 through allosteric interactions, and the effect of mutations depends largely on their
29 surrounding environment.

30 The identification of functional sites on the proteins such as binding surfaces by
31 using evolutionary tracing of the conserved residues in the homologous sequences
32 and structures has been recently investigated.^{20,21} There are a few studies where
33 the conservation of residues was assessed qualitatively using multiple alignments
34 and phylogenetic trees to map evolutionary information onto data sets of protein
35 structures.^{16,22} Recently Ofran and Rost²³ have classified the interface type into
36 six classes and noted the composition of amino acids sequence in each of these
37 classes. In another study, Caffrey *et al.*²⁴ have analyzed sets of heterodimer and
38 homodimer proteins separately, and found that interface regions are only slightly
39 more conserved than other regions of the protein structures, similar to some of the
40 earlier observations.²⁵ Caffrey *et al.*²⁴ did not find any significant differences in the
41 conservation patterns in the hetero-, homo-dimer transient and non-transient com-
42 plex structures. This indicates that residue conservation is similar in all kinds of
43 functionally involved interaction sites. Hence, the question still stands of whether

1 evolutionary processes constrain mutations on the interface of protein-protein
2 complexes.

3 In order to properly evaluate and quantify the residue conservation at the inter-
4 face region of protein structures, here we present results from our analysis of a set
5 of well-resolved protein complex structures. We also studied an additional 38 non-
6 antigen-antibody complex structures and 19 antigen-antibody structures separately
7 from the benchmark of Chen *et al.*²⁶ to validate our findings.

8 Our results clearly indicate that the number surface density of highly conserved
9 residues is significantly higher at the interface region of protein complexes. However,
10 the average conservation index of residues in the interaction patch is only slightly
11 higher than that in any other part of the protein surface. The analysis shows that
12 the actual number of highly conserved residues per interaction site is a more use-
13 ful indicator for predicting protein-protein binding sites than the average value of
14 conservation index of patches.³²

15 2. Materials and Methods

16 We have selected a set of non-homologous hetero complex structures defined by
17 X-ray crystallography at 2.0 Å or better resolution from the SPIN_PP data base
18 (<http://trantor.bioc.columbia.edu/cgi-bin/SPIN/>). From these structures we have
19 used only the dimer complexes which have at least 1000 Å² interface area between
20 the protein dimers (i.e. a total of 2000 Å² on both protomers of complex). If the
21 complex is a polymer of more than two proteins we have isolated a single particular
22 dimer from the complex and used that for analysis. There were 28 dimer complexes
23 in the protein data bank²⁷ (PDB) which satisfy the above criteria and have a
24 significant number of homologous sequences of 8 or more available in the UniProt
25 sequence database.³³ These complexes, used in our analysis, are given in Table 1. We
26 have further validated our findings on a separate data set of non-antigen-antibody
27 benchmark structures (Table 2) suggested by Chen *et al.*²⁶ The antigen-antibody
28 complex structures of the benchmark (Table 3) were studied separately.

29 2.1. Homologous sequences

30 We obtained homologous sequences of each polypeptide of the complex from
31 the UniProt database, an annotated non-redundant protein sequence data base
32 (A non-redundant swissprot + TrEMBL + TrEMBLnew), using the FASTA3
33 (<http://www.ebi.ac.uk/fasta33/>) sequence similarity search tool at the European
34 Bioinformatics Institute. Final homologous sequence alignments were obtained
35 using the MVIEW tool at the same site. Homologous sequences with less than 30%
36 gaps in the sequence and greater than 35% sequence identity to the parent sequence
37 were used for analysis. If the evolutionary distance (described below) between any
38 two sequences is less than 5% then we randomly remove one of the sequences from
39 the homolog set. The remaining sequences were used for calculating the residue con-
40 servation index (described below). This ensures that clean homologous sequences

4 *B. V. B. Reddy & Y. N. Kaznessis*

Table 1. Best resolved heterocomplexes that are used for analysis. Protein Data Bank Identification code (ID), X-ray resolution of the structure (Å), Protein Chain identifier in PDB (Ch), Length of the sequences (Len). Number of clean sequence homologues (Hom) obtained from the UniProt database.

ID	Å	Ch	Len	Hom	Ch	Len	Hom	ID	Res Å	Ch	Len	Hom	Ch	Len	Hom
1aok	2.0	A	122	189	B	122	192	1mee	2.0	A	275	205	I	64	10
1apy	2.0	A	161	15	B	141	18	1opg	2.0	L	214	24	H	227	53
1aqq	1.84	H	225	50	L	215	26	1osp	1.95	L	214	24	H	218	52
1cpc	1.66	A	162	44	B	172	42	1phn	1.65	A	162	33	B	171	76
1dan	2.0	L	132	39	H	254	275	1sct	2.0	A	149	15	B	150	12
1dan	2.0	L	132	39	U	116	9	1slu	1.8	A	142	6 ^a	B	216	56
1dan	2.0	T	75	9	U	116	9	1spb	2.0	P	71	9	S	264	132
1dvf	1.9	C	107	239	D	120	214	1spg	1.95	A	143	143	B	147	179
1fgv	1.9	L	107	221	H	120	209	1vge	2.0	L	214	25	H	225	57
1fle	1.9	E	240	277	I	47	53	1yec	1.9	L	214	18	H	222	55
1ftr	1.85	L	219	23	H	216	51	2fbj	1.95	L	213	23	H	220	52
1hds	1.98	A	141	232	B	145	111	2sic	1.6	E	275	205	I	107	24
1hyx	1.8	L	216	22	H	224	54	7fab	2.0	L	204	28	H	209	54
1kel	1.9	L	217	24	H	218	55	8fab	1.8	A	206	25	B	215	38

^aIndicates chains not used in calculation of Residue Conservation Index because of the low number of homologous sequences.

Table 2. Protein complexes used as a benchmark for docking. Protein Data Bank Identification code (ID), X-ray resolution of the structure (Å), Protein Chain identifier in PDB (Ch), Length of the sequences (Len). Number of clean sequence homologues (Hom) obtained from the UniProt database.

ID	Å	Ch	Len	Hom	Ch	Len	Hom	ID	RÅ	Ch	Len	Hom	Ch	Len	Hom
1a0o	2.95	A	128	101	B	134	4 ^a	1ppe	1.8	E	223	325	I	27	21
1acb	2.0	E	245	300	I	68	23	1spb	2.0	P	78	11	S	266	188
1ahw	3.0	E	214	159	F	219	8	1stf	2.37	E	212	320	I	98	27
1atn	2.8	A	373	66	D	258	24	1tab	2.3	E	223	325	I	82	51
1avw	1.75	A	223	333	B	171	39	1tgs	1.8	I	56	74	Z	229	325
1avz	3.0	B	151	56	C	57	160	1udi	2.7	E	244	121	I	223	126
1brc	2.5	E	223	316	I	56	167	1ugh	1.9	E	223	126	I	76	1 ^a
1cgi	2.3	E	245	300	I	56	59	1wq1	2.5	G	334	3 ^a	R	166	159
1cho	1.8	E	241	292	I	55	74	2btf	2.55	A	375	61	P	140	19
1cse	1.2	E	274	197	I	70	24	2kai	2.5	A	80	293	B	152	271
1dfj	2.5	E	124	143	I	457	8						I	58	140
1efu	2.5	A	385	171	B	282	99	2mta	2.4	A	105	17	L	125	4 ^a
1fin	2.3	A	298	261	B	260	148						H	373	5 ^a
1fq1	2.0	A	212	6 ^a	B	298	261	2pcc	2.3	A	296	48	B	108	125
1fss	3.0	A	537	110	B	60	52	2ptc	1.9	E	223	325	I	58	140
1gla	2.6	F	168	70	G	501	118	2sic	1.8	E	275	204	I	107	24
1got	2.0	A	350	192	B	340	43	2sni	2.1	E	275	207	I	83	32
					G	73	12	2tec	1.98	E	279	175	I	68	23
1kkl	2.8	A	205	34	C	205	34	3hhr	2.8	A	190	66	B	203	34
					H	100	55	4htc	2.3	H	259	240	I	65	9
1l0y	2.5	A	236	10	B	221	22						L	35	12
1mah	3.2	A	543	129	F	60	52								

^aIndicates chains not used in calculation of Residue Conservation Index because of low number of homologous sequences.

Table 3. Antigen-antibody complex structures. Protein Data Bank Identification code (ID), X-ray resolution of the structure (Å), Protein Chain identifier in PDB (Ch), Length of the sequences (Len). Number of clean sequence homologues (Hom) obtained from the UniProt database.

ID	Å	Ch	Len	Hom	Ch	Len	Hom	Ch	Len	Hom
1ahw	3.0	D	214	46	E	214	159	F	219	8
1bql	2.6	H	215	155	L	212	48	Y	129	94
1bvk	2.7	D	108	343	E	117	457	F	129	93
1dqj	2.0	A	214	47	B	210	137	C	129	93
1eo8	2.8	H	217	161	L	210	47	A	328	11
1fbi	3.0	H	221	169	L	214	46	X	129	84
1iai	2.9	H	219	165	L	214	46	I	218	161
1iai	2.9	—	—	—	—	—	—	M	215	47
1jhl	2.4	H	116	459	L	108	326	A	129	91
1kxq	1.6	D	496	92	E	120	495	—	—	—
1kxt	2.0	A	496	91	B	127	492	—	—	—
1kxv	1.6	A	496	92	C	121	493	—	—	—
1mel	2.5	B	148	484	M	129	93	—	—	—
1mlc	2.1	A	214	48	B	218	152	E	129	93
1nca	2.5	H	221	167	L	214	48	N	389	21
1nmb	2.5	H	122	457	L	109	288	N	470	56
1qfu	2.8	H	223	168	L	217	45	A	328	11
1wej	1.8	H	223	156	L	214	47	F	105	139
2jel	2.5	H	218	160	L	217	49	P	85	53
2vir	3.25	A	210	52	B	221	165	C	282	14

^aIndicates chains not used in calculation of Residue Conservation Index because of low number of homologous sequences.

1 with sufficient divergence are used with appropriate weight as per their evolution-
 2 ary distances. We have only used the complexes which have at least 8 homologous
 3 structures available after all the filters. The main assumption of the analysis is that
 4 the chosen homologous proteins have similar structure and roughly have the same
 5 binding site in the sequence.

2.2. Evolutionary distance

7 Evolutionary distance among the sequences is calculated using Eq. (1).

$$ED_{ij} = \left[\left(\left(1 - \frac{S_{ij}}{S_{ii}} \right) + \left(1 - \frac{S_{ij}}{S_{jj}} \right) \right) / 2 \right] \times 100 \quad (1)$$

9 A similarity score S_{ii} for sequence i is calculated by summing up the iden-
 10 tical substitution (diagonal values of substitution matrix, Gonnet *et al.*²⁹). Simi-
 11 larly, the S_{jj} score is calculated for sequence j . A similarity score S_{ij} between the
 12 sequences i and j is calculated using substitution matrix values of corresponding
 13 aligned residues between the two sequences.

6 *B. V. B. Reddy & Y. N. Kaznessis*

1 **2.3. Conservation index of residue position**

3 As described above, evolutionary distances between the reference sequence and its
 4 homologues were used to calculate the residue conservation index, CI_l , for each
 5 position l using the amino acid substitution matrix, which is similar to the amino
 6 acid variability or conservation used by Sander, Schneider,²⁸ Valdar and Thornton.⁸
 7 Conservation Index (CI_l) is a weighted sum of pairwise similarities between all
 8 residues present at the position. The CI_l value is calculated using Eq. (2) in a
 9 given alignment and takes a value in the range $[0,1]$.

$$9 \quad CI_l = \frac{\sum_i^N \sum_{j>i}^N ED(s_i) \times ED(s_j) \times Mut(s_i(l), s_j(l))}{\sum_i^N \sum_{j>i}^N ED(s_i) \times ED(s_j)} \quad (2)$$

11 where N is the number of homologous sequences in the alignment; $s_i(l)$ and $s_j(l)$
 12 are the amino acids at the alignment position l of sequences s_i and s_j respectively;
 13 $ED(s_i)$ and $ED(s_j)$ are the average evolutionary distances of $s(i)$ and $s(j)$ from the
 14 remaining homologues. $Mut(a,b)$ measures the similarity among the amino acids a
 15 and b as derived from amino acid substitution matrix $M(a,b)$ and is defined as

$$15 \quad Mut(a,b) = \frac{M(a,b) - M(a,b)_{\text{low}}}{M(a,b)_{\text{max}} - M(a,b)_{\text{low}}} \quad (3)$$

17 where a, b are the pairs of amino acids at a given alignment position l . $M(a,b)_{\text{low}}$ is
 18 the lowest value in the substitution matrix (-5 in the Gonnet *et al.*, 1992 matrix)
 19 and $M(a,b)_{\text{max}}$ is the maximum value among all the possible substitution pairs in
 20 that position. Thus the $Mut(a,b)$ takes a value in the range $[0,1]$. This method of cal-
 21 culating the conservation index ensures that the homologous sequences are appro-
 22 priately weighed depending on their evolutionary distance between the sequences
 23 (the higher the ED the greater the weight).

23 **2.4. Interfacial amino acids in protein complexes**

25 The solvent accessible surface area (SASA) of individual amino acids is calcu-
 26 lated using the method of Richmond and Richards³⁰ as implemented by Sali
 27 and Blundell.³¹ Residue X is said to have 100% solvent accessibility in the Gly-
 28 X-Gly form of the linear tri-peptide and other percentages of accessibilities are
 29 referred with reference to this value. We have not used the absolute value in square
 30 angstroms since the SASA of each amino acid depends on the size of the residue.
 31 SASA values were used to identify surface residues, buried residues and the inter-
 32 face residues in the complex structures as follows. We have calculated SASA of each
 33 of the residues in each protomer in the presence and in the absence of complexation
 and the difference in SASAs is taken as the interface contact area of the residues
 in the interfacial region.

2.5. Average conservation index of protein surface patches

Around each surface residue we identified neighboring surface residues whose C_β atoms (C_α in the case of Glycine) fall within a sphere of a given radius, and defined the group as a surface patch around that residue. We calculated an average conservation index (ACI) of positions in each patch of 10 Å radii. The number of interfacial residues present in the patch and the ACI values are given in Table 4.

3. Results and Discussion

We have investigated the conservation of interfacial residues in the 28 X-ray crystallography defined hetero dimer protein complex structures, which are unique and resolved at 2.0 Å or at a better resolution (Table 1). These structures consist of non-antigen-antibody interactions. A total of 12 220 amino acid positions are present in these proteins, 9095 positions have residues with greater than 10% solvent accessible area and 2060 amino acid positions are at the interface with at least 10% of the area buried in the interface region. The comparison between the amino acid composition of interface residues and the surface residues is shown in Fig. 1. It can be seen from the figure that, as expected, sites with hydrophilic amino acids such as K, D, E, N, Q are relatively more frequent on the protein surfaces. However aromatic amino acid positions, F, Y, W are more frequent in the interface region of the protein complexes. Positions of L and P amino acids are also relatively more frequent in the interface region of the protein complexes.

As described in the Methods section we computed the conservation index (CI) for each position of the sequence using the set of their homologue sequences of each of the selected heterodimers. We have used the amino acid substitution matrix ($M(a, b)$)²⁹ and the evolutionary distance (ED) of the sequence to calculate CI values. The higher the CI value the more the residue (or the property of the residue)

Table 4. Comparison of highly conserved positions in the interfacial and non-interfacial surface region of the protein dimer structures. The number of residue occurrence positions (fraction) in each group of conservation indices is given. Ratio of the fraction of residues in the interface region compared to non-interface region is also given.

Residue Solvent Accessibility	Interfacial Contact Area	Conservation Index of Group Intervals			
		< 0.44	0.44–0.61	0.61–0.85	> 0.85
All residues		3086(.252)	3052(.249)	3073(.251)	3009(.246)
> 10%	< 10%	2059(.293)	2057(.292)	1771(.252)	1148(.163)
	≥ 10%	561(.272)	466(.226)	493(.239)	540(.262)
	Ratio	0.930	0.774	0.951	1.606
> 20%	< 20%	1926(.302)	1937(.303)	1613(.253)	909(.142)
	> 20%	388(.269)	316(.219)	341(.236)	399(.276)
	Ratio:	0.891	0.721	0.935	1.941
	< 10%	1791(.301)	1821(.306)	1504(.253)	835(.140)
	> 10%	523(.278)	432(.230)	450(.240)	473(.252)
	Ratio:	0.925	0.752	0.948	1.795

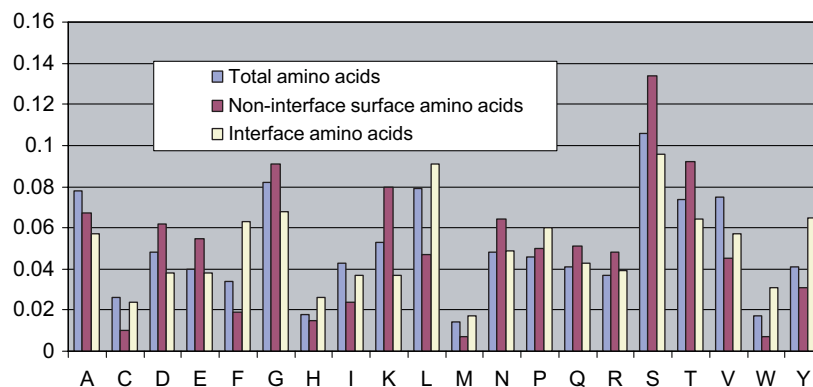
8 *B. V. B. Reddy & Y. N. Kaznessis*

Fig. 1. Fraction of amino acid occurrences in interfacial regions and non-interfacial regions of protein complex surfaces compared with the total occurrences in the protein complexes.

1 is conserved at that position. Based on the CI values, which vary between 0 and 1,
 2 we have divided all the sequence positions into four groups in such a way that the
 3 residues are distributed equally among the four group intervals: group 1 positions
 4 have CI values less than or equal to 0.44, group 2 positions have CI values between
 5 0.44–0.61, group 3 positions have values between 0.61–0.85 and group 4 positions
 6 have values greater than 0.85 (Table 4). In each group there are approximately 25%
 7 of the total residue positions.

3.1. Density of conserved positions on the interface

9 In Table 4 the distribution is shown for amino acid positions in each conserva-
 10 tion class. We have shown data for two different cutoffs of 10% and 20% solvent
 11 accessibility to differentiate between surface residues, interface residues and buried
 12 residues.

13 Among the positions with residue surface accessibility larger than 10% and
 14 interfacial contact area of less than 10%, the percentage of highly conserved residue
 15 positions (group 4) decreases to 16%. However, the percentage of these group 4
 16 positions remained at about 26% for the interfacial residues.

17 Focusing on the ratio between the fractions of conserved position occurrences
 18 in the interface positions and the non-interface positions, it is less than one in the
 19 first three low conservation index groups, whereas in the high conservation group 4
 20 the ratio is 1.6.

21 Similar trends are observed when the limit of the solvent accessibility is increased
 22 to 20% of the residue area to define the surface, interface and buried residues. The
 23 ratio of highly conserved residue positions is substantially increased in the interfacial
 24 region. These observations clearly show that highly conserved positions are found
 25 more frequently in the interface region of the protein complexes.

1 From the values in Table 4 one can calculate the number surface density of
 2 group 4 positions on protein surfaces. On non-interfacial surfaces, with residue
 3 accessibility $>10\%$ and interfacial contact area $<10\%$, approximately one out of
 4 6 positions is a group 4 (highly conserved) position. On the interfacial surfaces,
 5 with residue accessibility $>10\%$ and interfacial contact area $\geq 10\%$, approximately
 6 one out of 4 positions is a well conserved group 4 position. In the case of $>20\%$
 7 accessible residues with $<20\%$ of interfacial area, the density of the conserved
 8 group 4 positions on the non-interfacial surface is one in 7 positions, whereas in
 9 the interfacial surface, $>20\%$ accessible residues with $\geq 20\%$ of interfacial area, it is
 10 one in 3.6 positions. This clearly shows that the number surface density of highly
 11 conserved residues is significantly higher in the interface region compared with the
 non-interface regions of protein structures.

13 3.2. Average conservation index of interfacial residues

14 We have also studied patches of the surface residues in a given radius around each
 15 surface residue to investigate whether the average conservation index is higher for
 16 the interfacial patches than for the non-interfacial ones. In Table 5, patches are
 17 presented based on the number of interface residues (IR) they contain. We have
 18 investigated patches with radius of 10 \AA .

19 The average number of surface residues in the patch, the total occurrence of such
 20 patches and the average conservation index of the residues present in the patch are
 21 calculated and presented in Table 5. A residue is considered to be present on the
 surface if at least 10% of the surface is accessible to the solvent.

Table 5. Average conservation of surface residues in each patch of 10 \AA radius. Interface residues (IR), average number of surface residues in each patch (ANSR), total number of patches (NC) and the average conservation index (ACI) of residue positions in the corresponding patches are given along with their standard deviations. An amino acid accessible area and interface residue contact area of 10% or greater is used to define the surface residues.

IR	ANSR	NC	ACI
0	10.37 ± 2.39	3412	0.601 ± 0.147
1	10.52 ± 2.55	619	0.612 ± 0.146
2	10.58 ± 2.62	530	0.602 ± 0.151
3	10.49 ± 2.56	496	0.616 ± 0.157
4	10.44 ± 2.47	518	0.620 ± 0.153
5	10.80 ± 2.43	482	0.618 ± 0.156
6	11.04 ± 2.40	345	0.628 ± 0.162
7	11.68 ± 2.28	249	0.635 ± 0.155
8	12.18 ± 2.22	146	0.660 ± 0.160
≥ 9	13.19 ± 2.04	184	0.649 ± 0.139

1 In this analysis, the patches with 0 to 3 interface residues have an average
conservation index (ACI) of about 0.60 and the patches with more interface residues
3 (>4) have about 0.63 to 0.65 ACI. Considering the standard errors, it is clear that
the ACI of patches with higher number of interface residues is not significantly high
5 in order to distinguish the interfacial patches from the non-interfacial patches. This
analysis indicates that the average residue conservation is only slightly higher in
7 the interface region of the protein interaction sites, which is in agreement with the
work of Caffrey *et al.* The ACI values are lowered! because in each interfacial patch
9 there is considerable number of surface residues with low conservation index.

According to our previous calculations there is on average one highly conserved
11 position in every 6.12 positions in the non-interfacial regions and in every 3.81
highly conserved positions in the interfacial region. On average an interface patch
13 of 20 residues has 3.27 highly conserved positions in the non-interface region and
5.25 highly conserved positions in the interface region. There is a difference of
15 1.98 highly conserved positions in the interface region. When calculating ACI the
increase in CI value averaging over 20 positions therefore will not show any signifi-
17 cant increase (Table 5).

This analysis clearly shows that the average conservation index will not be
19 useful in predicting functional interaction sites on protein surfaces. Nonetheless,
the results indicate that for interactions which are functionally very important, the
21 number of highly conserved residues in the interfacial region may help us to identify
the putative interaction sites on the given protein structures.³²

23 Interestingly, in Table 5 we also find an increase in the average number of surface
residues per patch in the patches that have higher number of interface residues. In
25 other words the interface patches are found to have slightly more compact packing
compared with non-interface patches. This could be because the interface packing
27 can be considered similar to the packing in the core region of the protein, with
more hydrophobic residues involved in tight packing. In contrast, the non-interfacial
29 surface region may be more flexible, with appropriate gaps between the polar surface
atoms to allow for solvent molecules to interact more freely with the polar atoms
31 on the surface.

3.3. *Validation on benchmark complex structures*

33 The results above were obtained using a smaller group of well resolved pre-
dominantly non-transient hetero complexes. In order to validate our results we
35 have used an additional set of benchmark structures suggested by Chen *et al.*,²⁶
employed recently on several protein docking studies. We have considered non-
37 antigen-antibody complexes (Table 2) and antigen-antibody complexes (Table 3)
separately in this analysis. It can be seen from Table 6 that we have used the same
39 four class intervals of conservation index values and computed the occurrence of
surface residue positions with their respective conservation indices. In the case of

Table 6. Conservation of positions in the interfacial and non-interfacial surface region of the benchmark structures. A loss in solvent accessibility of 10% or greater upon complex formation is used as a criterion to define interfacial (IN) and otherwise non-interfacial (NIN) residue. The ratios of the fractional occurrences are also given in a separate row for each group of complex structures.

	Conservation Index of Group Intervals			
	< 0.44	0.44–0.61	0.61–0.85	> 0.85
(i) Non antigen-antibody complexes				
NIN	3266(0.42)	2058(0.26)	1497(0.19)	1009(0.13)
IN	405(0.35)	253(0.22)	259(0.23)	225(0.20)
Ratio:	0.850	0.843	1.186	1.529
(ii) Antigen-antibody complexes				
NIN	1128(0.16)	2154(0.31)	2169(0.31)	1440(0.21)
IN	282(0.49)	147(0.25)	106(0.18)	43(0.07)
Ratio:	2.980	0.814	0.583	0.356
(iii) Only antibody regions				
NIN	860(0.17)	1729(0.34)	1700(0.34)	761(0.15)
IN	196(0.61)	73(0.23)	43(0.13)	9(0.03)
Ratio:	3.585	0.664	0.398	0.186
(iv) Only antigen regions				
NIN	268(0.15)	425(0.23)	469(0.25)	679(0.37)
IN	86(0.33)	74(0.29)	63(0.25)	34(0.13)
Ratio:	2.299	1.247	0.962	0.359

1 non-antigen-antibody complex structures, similar to the results in Table 4, the frac-
 2 tion of highly conserved residue positions is higher in the interface regions compared
 3 with the non-interface surface regions. In the case of antigen-antibody complexes
 4 the reverse pattern is observed with a small number of conserved positions in the
 5 interacting regions. Since antibodies are generated with variable amino acids as
 6 a defense mechanism to interact with different antigens, we expect the interact-
 7 ing regions to be variable and these are not the functional part of the evolved
 8 interactions. Interestingly, antigen interacting regions also have a higher number
 9 of non-conserved residue positions in the interacting region than the rest of the
 10 protein surface (Table 6).

11 The calculated conservation index values depend on the number of available
 12 homologous sequences and the evolutionary distance among them and hence, we
 13 also calculated a relative conservation index of positions. We arranged all the sur-
 14 face residues in a descending order according to their conservation index and com-
 15 puted the ratio of their fractional occurrence in the interfacial versus non-interfacial
 16 regions in each complex. We thus obtained the average value and the correspond-
 17 ing standard deviations (Table 7). For about 82% of the complexes the ratio is
 18 higher than 1.0 with average values of 1.28 or greater depending on how interfacial
 19 residues are defined (either > 10% or > 20% contact area). In Table 7 the standard
 20 deviations are high for the top 10% of the highly conserved positions, but decreases
 21 for the top 20% or 30% of the conserved positions, while the average ratio remains
 at about 1.3.

12 *B. V. B. Reddy & Y. N. Kaznessis*

Table 7. Ratio of highly conserved surface residues in interface region versus non-interface regions. The positions of amino acids with residues $\geq 10\%$ or $\geq 20\%$ of their area in contact are defined as interface residue positions.

Conserved Positions in the Structures	Ratio of Highly Conserved Residue Positions in the Interface Region and in Non-interface Region			
	Complexes of Table 1 % of Area in Contact		Complexes of Table 2 % of Area in Contact	
	$\geq 10\%$	$\geq 20\%$	$\geq 10\%$	$\geq 20\%$
Top 10%	1.68 ± 0.90	2.03 ± 1.03	1.76 ± 1.41	2.00 ± 1.74
Top 20%	1.39 ± 0.48	1.59 ± 0.49	1.43 ± 0.87	1.61 ± 1.07
Top 30%	1.28 ± 0.34	1.40 ± 0.40	1.22 ± 0.65	1.33 ± 0.72

1 Among the top 30% of the highly conserved surface positions (20% contact
 2 area), interface positions have 40% more conserved positions than the non-interface
 3 ones. We feel that this difference is significant enough for assisting the prediction
 4 of interface positions on the protein surface.³²

5 It should be noted that we have assumed that the surface regions that do not
 6 participate in each of the specific complex that we examined are altogether non-
 7 interacting sites, which may not be true *in vivo*. Some of these surface regions might
 8 actually be involved in transient complexation with other protein molecules, not
 9 included in the crystal structures. Hence, in our analysis these sites might appear
 10 as false positives (may have significantly large number of conserved positions but
 11 non-interacting) although in actuality some of them may be binding sites.

3.4. Concluding remarks

13 In summary, we observe that the number surface density of highly conserved residue
 14 positions is higher in the interfacial region of protein complexes compared with the
 15 non-interfacial regions. As demonstrated in the subsequent work³² the difference is
 16 observed to be significant enough for assisting the prediction of interface positions
 17 on the protein surface. On the other hand, the average conservation index of residues
 18 in the interface regions is not significant due to the relatively small number of
 19 highly conserved positions compared with the total positions with low conservation
 20 indices at an interface. The analysis further shows that the interacting regions in an
 21 antibody-antigen complex occur mostly through non-conserved positions for both
 22 the antibody and the antigen.

23 Acknowledgments

24 We gratefully acknowledge the support from the Digital Technology Center, Uni-
 25 versity of Minnesota. This work was also partially supported by the University of
 26 Minnesota Bioinformatics Institute and the American Chemical Society Petroleum
 27 Research Fund Grant (Award no. G7-38758).

1 **References**

- 3 1. Lijnzaad P, Argos P, Hydrophobic patches on protein subunit interfaces: characteristics and prediction, *Proteins* **28**:333–343 (1997).
- 5 2. Bogan AA, KS Thorn, Anatomy of hot spots in protein interfaces, *J Mol Biol* **280**:1–9 (1998).
- 7 3. Chothia C, Janin J, Principles of protein-protein recognition, *Nature* **255**:705–708 (1975).
- 9 4. Argos P, An investigation of protein subunit and domain interfaces, *Protein Eng* **2**:101–113 (1988).
- 11 5. Janin J, Chothia C, The structure of protein-protein recognition sites, *J Biol Chem* **265**:16027–16030 (1990).
- 13 6. Jones S, Thornton JM, Protein-protein interactions: a review of protein dimer structures, *Prog Biophys Mol Biol* **63**:31–65 (1995).
- 15 7. Jones S, Thornton JM, Analysis of protein-protein interaction sites using surface patches, *J Mol Biol* **272**:121–132 (1997).
- 17 8. Valdar WS, Thornton JM, Protein-protein interfaces: analysis of amino acid conservation in homodimers, *Proteins* **42**:108–124 (2001).
- 19 9. Archakov AI, Govorun VM, Dubanov *et al.*, Protein-protein interactions as a target for drugs in proteomics AV, *Proteomics* **3**:380–391 (2003).
- 21 10. Young L, Jernigan RL, Covell DG, A role for surface hydrophobicity in protein-protein recognition, *Protein Sci* **3**:717–729 (1994).
- 23 11. Jones S, Thornton JM, Prediction of protein-protein interaction sites using patch analysis, *J Mol Biol* **272**:133–143 (1997).
- 25 12. Livingstone CD, Barton GJ, Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation, *Comput Appl Biosci* **9**:745–756 (1993).
- 27 13. Zvelebil MJ, Barton GJ, Taylor WR, Sternberg MJ, Prediction of protein secondary structure and active sites using the alignment of homologous sequences, *J Mol Biol* **195**:957–961 (1987).
- 29 14. Shakhnovich E, Abkevich V, Ptitsyn O, Conserved residues and the mechanism of protein folding, *Nature* **379**:96–98 (1996).
- 31 15. Reddy BVB, Li WW, Shindyalov IN, Bourne PE, Conserved key amino acid positions (CKAAPs) derived from the analysis of common substructures in proteins, *Proteins*. **42**:148–163 (2001).
- 33 16. Lichtarge O, Bourne HR, Cohen FE, An evolutionary trace method defines binding surfaces common to protein families, *J Mol Biol* **257**:342–358 (1996).
- 35 17. Teichmann SA, Principles of protein-protein interactions, *Bioinformatics Suppl* **2**:S249 (2002).
- 37 18. Apweiler R, Attwood TK, Bairoch *et al.*, The InterPro database, an integrated documentation resource for protein families, domains and functional sites A, *Nucleic Acids Res* **29**:37–40 (2001).
- 39 19. Lockless SW, Ranganathan R, Evolutionarily conserved pathways of energetic connectivity in protein families, *Science* **286**:295–299, (1999).
- 43 20. Lichtarge O, Sowa ME, Evolutionary predictions of binding surfaces and interactions, *Curr Opin Struct Biol* **12**:21–27 (2002).
- 45 21. Glaser F, Pupko T, Paz *et al.*, ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information I, *Bioinformatics* **19**:163–164 (2003).
- 47 22. Mihalek I, Res I, Lichtarge O, A family of evolution — entropy hybrid methods for ranking protein residues by importance, *J Mol Biol* **336**:1265–1282 (2004).

14 B. V. B. Reddy & Y. N. Kaznessis

- 1 23. Ofraan Y, Rost B, Analysing six types of protein-protein interfaces, *J Mol Biol*
325:377–87 (2003).
- 3 24. Caffrey DR, Somaroo S, Hughes JD *et al.*, Are protein-protein interfaces more con-
 served in sequence than the rest of the protein surface, *Protein Sci* **13**:190–202 (2004).
- 5 25. Grishin NV, Phillips MA, The subunit interface of oligomer enzymes are conserved
 to a similar extent to the overall protein sequences, *Protein Sci* **3**:2455–2458, (1994).
- 7 26. Chen R, Mintseris J, Janin J, Weng Z, A protein-protein docking benchmark, *Proteins:
 Str Fun. Genetics* **52**:88–91 (2003).
- 9 27. Berman HM, Westbrook J, Feng *et al.*, Protein data bank Z, *Nucleic Acids Res*
28:235–242 (2000).
- 11 28. Sander C, Schneider R, Database of homology-derived protein structures and the
 structural meaning of sequence alignment, *Proteins* **9**:56–68 (1991).
- 13 29. Gonnet GH, Cohen MA, Benner SA, Exhaustive matching of the entire protein
 sequence database, *Science* **256**:1443–1445 (1992).
- 15 30. Richmond TJ, Richards FM, Packing of alpha-helices: geometrical constraints and
 contact areas, *J Mol Biol* **119**:537–555 (1978).
- 17 31. Sali A, Blundell TL, Definition of general topological equivalence in protein structures.
 A procedure involving comparison of properties and relationships through simulated
 annealing and dynamic programming, *J Mol Biol* **212**:403–428 (1990).
- 19 32. Duan Y, Reddy BVB, Kaznessis Y, Physicochemical and residue conservation calcu-
 lations to improve ranking of protein-protein docking solutions, *Protein Sci* in press.
- 21 33. Bairoch A, Apweiler R, Wu *et al.*, The universal protein resource (UniProt) CH, *Nucl*
 23 *Acids Res* **33**:D154–D159.



Boojala V. B. Reddy received his M.Sc. in Life Sciences from
 Jawaharlal Nehru University, New Delhi, India and his Ph.D.
 in Life Sciences from Centre for Cellular and Molecular Biology
 and University of Hyderabad, India and post doctoral training
 from Birkbeck College, University of London. He is author of
 several computational methods on DNA and protein sequence
 and structural data analysis modeling and structure prediction
 studies.



Yiannis N. Kaznessis is currently on Assistant Professor in
 Chemical Engineering and Materials Science and Director of the
 University of Minnesota Bioinformatics Summer Institute. He
 received his Diploma in Chemical Engineering at the Aristotle
 University of Thessaloniki, Greece and his Ph.D. at the Depart-
 ment of Chemical Engineering at the University of Notre Dame.
 He received postdoctoral training at the University of Michigan
 and at Pfizer Global Research and Development.