Bioinformation by Biomedical Informatics Publishing Group

open access

www.bioinformation.net

Hypothesis

Structural segments and residue propensities in protein-RNA interfaces: Comparison with protein-protein and protein-DNA complexes

Sumit Biswas[1], Mainak Guharoy[1], and Pinak Chakrabarti[1,*]

[1]Department of Biochemistry and Bioinformatics Centre, Bose Institute, Kolkata 700054, India;
Pinak Chakrabarti* - E-mail: pinak@boseinst.ernet.in; Phone: 91 33 2355 0256; Fax: 91 33 2355 3886; * Corresponding Author

received May 23, 2008; revised June 19, 2008; accepted July 07, 2008; published July 14, 2008

Abstract:
The interface of a protein molecule that is involved in binding another protein, DNA or RNA has been characterized in terms of the number of unique secondary structural segments (SSSs), made up of stretches of helix, strand and non-regular (NR) regions. On average 10-11 segments define the protein interface in protein-protein (PP) and protein-DNA (PD) complexes, while the number is higher (14) for protein-RNA (PR) complexes. While the length of helical segments in PP interaction increases with the interface area, this is not the case in PD and PR complexes. The propensities of residues to occur in the three types of secondary structural elements (SSEs) in the interface relative to the corresponding elements in the protein tertiary structures have been calculated. Arg, Lys, Asn, Tyr, His and Gln are preferred residues in PR complexes; in addition, Ser and Thr are also favoured in PD interfaces.

Keywords: protein-protein interactions; protein-DNA interactions; protein-RNA interactions; binding interface; protein secondary structure

Abbreviations: PP - protein-protein; PD - protein-DNA; PR - protein-RNA; SSE - secondary structural element; SSS - secondary structural segment

Background:
Characterization of protein-protein (PP), protein-DNA (PD) and protein-RNA (PR) interactions is essential for understanding the mechanisms of biological processes on a molecular level. Interactions are highly specific and any distortion may be deleterious to the cellular function. Various experimental techniques have been employed to identify the interactions **[1]**, with X-ray crystallography and NMR spectroscopy providing the most detailed view. The atomic coordinates of the complexes stored in the Protein Data Bank (PDB) **[2]** have been analyzed to derive information on the physicochemical features of the interface formed between the two components. PP interactions **[3-5]** have attracted the maximum attention. These can vary in strength – some are obligatory (permanent), as can be seen in the formation of the quaternary structures, while others are non-obligatory, in which the individual protomers exist independently in the stable form **[6]**, but the time scale of interaction can vary widely from $\sim 10^{-3}$ to 10^{3}s (transient to stable complexes, exemplified by electron transfer in redox proteins and antigen-antibody complexes, respectively). Studies in PD interactions have aimed at unravelling of the sequence specificity of nucleotide recognition **[7-11]**. In comparison PR interactions have been relatively fewer in number as data have been scarce till only recently **[12-14]**. Most of the complexes contain double-stranded DNA and the RNA is usually single stranded, though in a few cases depending on the sequence and length, it may fold into stem-loop structures including double helical segments. Akin to the non-obligatory PP complexes, PD and PR complexes are mostly transient, forming only when the protein encounters the nucleic acid, and exhibit a wide range of stability and lifetimes. With increase in our understanding of protein structure and interactions attempts are now geared towards synthetic biology for designing receptors for proteins and nucleic acids **[15]**. In this connection it is important to know what types of secondary structures are used in the interface and the residue usage vis-à-vis the rest of the protein structure. In this article these features are derived for PD and PR interfaces and compared to those observed in PP complexes **[16]**.

Methodology:
The list of 128 protein-DNA complexes used has been given in **[11]**. A search of PDB **[2]** (August, 2007) yielded 381 hits for the query "protein-RNA complex". The list of entries was culled using PISCES **[17]**, such that the maximum percentage identity was 25% and the resolution not worse than 3.0 Å. The minimum chain length for the protein part was kept at 40 and for RNA, at least 3 bases. For this non-redundant dataset of 50 protein-RNA complexes, the information on the biologically relevant assembly was obtained from the Nucleic Acid Database (NDB) **[18]** (since many PDB files have coordinates only for the crystallographic asymmetric unit, which may just contain a part of the whole molecule).

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

The protein secondary structural elements (SSEs) were assigned using DSSP **[19]**. Only three types of SSEs were considered. All helices (with DSSP notations 'H' and 'G') were included irrespective of their type, 'E' and 'B' constituted strands; turns ('T' and 'S') and the unclassified residues (with assignment ' ') together formed the nonregular (NR) region. Based on the presence of interface residues in distinct SSEs along the chain, the interface can be split into secondary structural segments (SSS) - a segment is specified by the span between the two extreme locations of the interface residues on that SSE (with or without intervening non-interface residues) **[16]**. The propensity $(P_i)^{SSE}$ of a residue *i* to occur in a given secondary structural element (SSE) was calculated by the following formula (1) under supplementary material.

Results and discussion:

Basic RNA-binding module and the interface area

Among the 50 RNA-binding proteins (Table 1, supplementary material) many are multimeric, each having distinct recognition sites which are structurally equivalent. Any one of them can be assumed to be the basic unit that gets repeated. We define this basic unit as one RNA-binding module (akin to what we have done for protein-DNA complexes **[11]**). The basic RNA-binding module that has been constructed can be repeated (by the application of simple symmetry operators) to generate the complete biological assembly. Thus for a homodimeric molecule (such as 1ooa), only one subunit interacting with the RNA was considered. In some other cases with more than two identical protein-RNA units (as in 2gic, where five identical protein chains bind symmetrically to five individual RNA strands), only one protein chain complexed with one RNA was considered. A considerable number (5) of the complexes in the dataset are coat proteins or nucleocapsids of viruses and bacteriophages. Basically, these are huge complexes (eg., 2fz2) formed by the application of a number of symmetry operators to a simple protein-RNA unit. For such complexes too, one subunit of the protein with one strand of the RNA was considered. 42 of the 50 complexes had the protein monomer binding to single-stranded RNA, and the rest to double-stranded RNA.

The interface area is given by the sum of the accessible surface area of the two isolated components minus that of the complex. This is the area that gets buried between the two components, which usually contribute almost equally **[4, 5]**. The average interface area in PR complexes is comparable to that observed in PD and PP complexes (Table 2 under supplementary material), though there is a larger variation around the mean. This is expected as the length of RNA located in the interface varies considerably (range: 3 to 37) among the different structures.

Secondary structural segments in the interface

Data presented in Table 2 (see supplementary material) indicate that there is not much distinction between the numbers of SSSs present in PP and PD interfaces, even when the value is normalized for a fixed size (1000 $Å^2$) of the interface. However, both these numbers are higher for PR interfaces. When the three SSSs (helix, strand and NR) are considered individually, the numbers are comparatively higher for PR than those in PD and PP interfaces. In contrast, the average lengths of the SSSs remain more or less the same in the three categories.

Variation of SSS length with interface size

The majority of the PP complexes have an interface with an area of 1600±400 $Å^2$ that has been termed as the standard size **[4]**. The variation of the segment lengths as a function of the interface size has also been addressed **[16]**. It was found that the average length of helix is doubled from ~4 when the area increases ten-fold from 500 $Å^2$; however, such changes were not observed for strand and NR segments. In comparison, in PD complexes (Figure 1a), a variation in the length of helical segments is not seen (the last bin is based on just single interface and is not considered) and a rather uniform length of 5.1-5.8 residues is observed, corresponding to ~1.5 turn of an α-helix interacting with the major groove of DNA. Interfaces have been classified as helical when the number of helical residues in the interface is more than 40% **[16]**. Considering strand and NR segments, in contrast to PP complexes, there are changes in PD complexes - the strand length decreases by about two fold and that of NR increases to the same extent. In PR complexes (Figure 1b), the length distribution, irrespective of segment type, is fairly uniform over the range of 500 to 3000 $Å^2$; but below 500 $Å^2$, the helical segments that are part of the interface contribute only two residues.

Secondary structure preferences of interface residues

Calculation of the propensities of residues to occur in a SSE in the PP interface relative to the same element in the overall protein tertiary structure indicated that Arg and the aromatic residues are observed more in all interface SSEs **[16]**. In PD complexes (Figure 2a), propensities > 1.5 are observed for the basic residues (Arg, Lys and His). Residues with hydroxyl groups (especially Ser and Thr) also have higher values. Residues with amide side-chains, Asn in particular, is found more in PD interfaces. Of the aromatic residues, Phe is less abundant. Gly, which is underrepresented in interface SSEs in PP complexes, is found more in PD complexes. It may be noted that unlike the PP complexes, the hydrophobic residues are unfavoured in PD interfaces, which are more polar in nature **[7, 8, 11]**. In PR complexes the highest propensities are observed in Arg, Asn and Lys (Figure 2b). Tyr, His and Gln are also preferred, but Ser and Thr, unlike in PD complexes, are not as favoured. Of interest is the fact that Asp, which is known to have a high propensity to be located in strands that form β-sheet across PP interfaces **[16]**, is also highly represented in strands in PR interfaces. This is in sharp contrast to the protein-DNA interfaces where Asp is poorly represented. Met located in NR segments is also preferred in PR complexes.

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

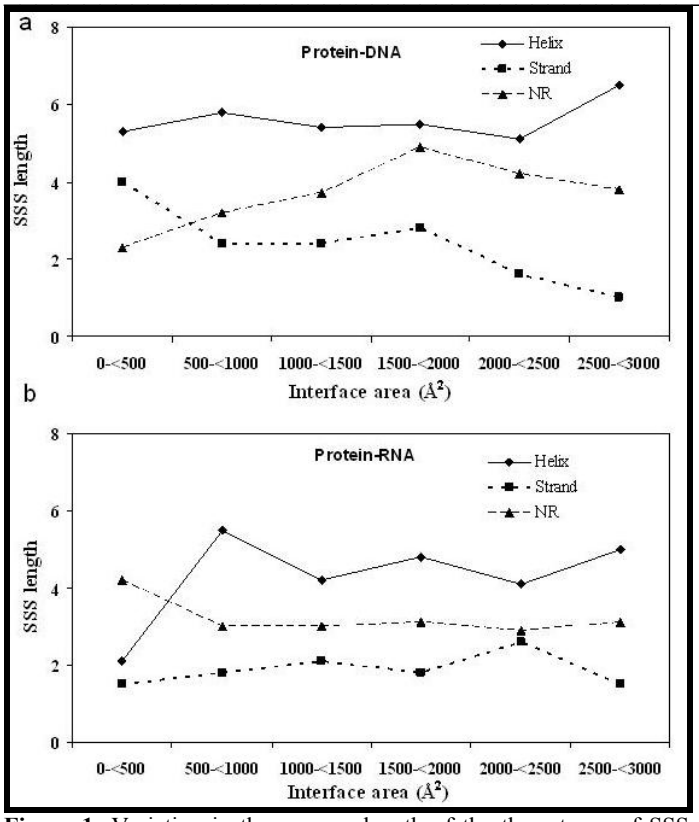

Figure 1: Variation in the average length of the three types of SSSs with the interface area contributed by the protein for (a) protein-DNA and (b) protein-RNA complexes. The protein interfaces have been grouped into bins of size 500 Å^2 and the average in each bin is plotted (the last bin in PD complexes has only one data point).

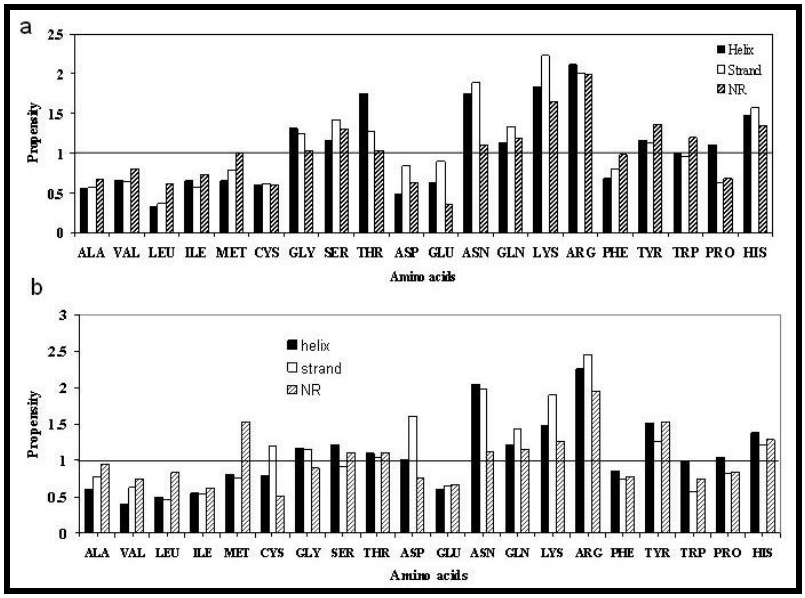

Figure 2: Propensities of residues to occur in a particular SSE in the interface relative to the same SSE in the tertiary structure in (a) protein-DNA and (b) protein-RNA complexes. A propensity of 1.0 indicates that the frequency of observing a particular residue in a given SSE in the interface is the same as that of the corresponding frequency for the entire tertiary structure; a value >1 (or <1) indicates higher (or lower) occurrence at the interface.

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Conclusion:

A non-redundant dataset of PR complexes has been created. This and a similar dataset of PD complexes **[11]** have been analyzed in terms of SSSs that constitute the protein part of the interface. PP complexes can bury a larger surface area by the use of longer helical segments **[16]**. However, in PR complexes the SSS length is rather invariant (Figure 1), but their number tends to increase regularly with the interface size (Figure 3). At any given size of the interface, the number of segments in PR complexes is usually more than that in PP complexes. In PD complexes the helices are of uniform length, but the strands get shorter with the concomitant increase in the length of NR, as the size of the interface increases. Relative to the tertiary structure the interface SSEs are depleted in Ala, Val, Ile and Leu in all types of complexes - interestingly these are the residues that have high α-helix or β-sheet propensities **[20]**. Compared to PP interfaces, aromatic residues are less favoured at the binding sites of nucleic acids. Preferred residues in PR complexes are Arg, Lys, Asn, Tyr, His and Gln; PD interfaces are also enriched in Ser and Thr. One feature that is common to both PP and PR interfaces is the presence of Asp in interface strands. The residue usage in a SSE in the interface relative to that in the overall tertiary structure would be useful in the design of structural motifs capable of interacting with another protein or nucleic acid.

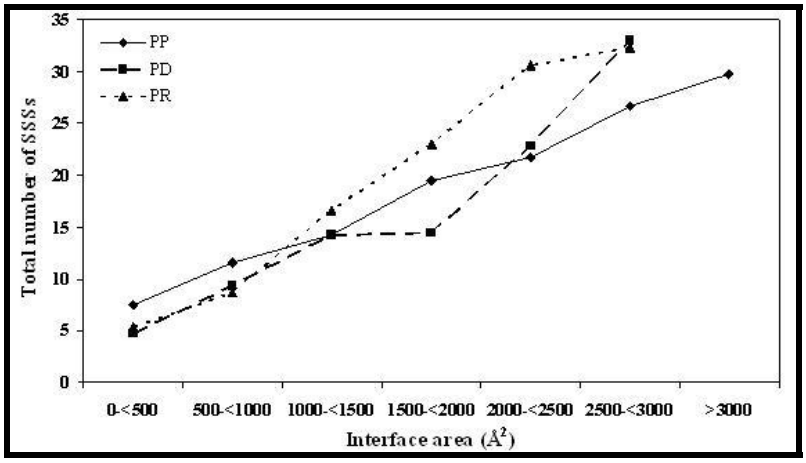

Figure 3: Variation in the average number of SSSs with the interface area (the details are in Figure 1 legend).

References:

[01] B. A. Shoemaker and A. R. Panchenko, *Plos. Comp. Biol.,* 3: 337 (2007) [PMID: 17397251]

[02] H. M. Berman *et al.*, *Nucleic Acids Res.,* 28: 235 (2000) [PMID: 10592235]

[03] S. Jones and J. M. Thornton, *Proc. Natl. Acad. Sci. USA,* 93: 13 (1996) [PMID: 8552589]

[04] L. Lo Conte, *J. Mol. Biol.,* 285: 2177 (1999) [PMID: 9925793]

[05] P. Chakrabarti and J. Janin, *Proteins,* 47: 334 (2002) [PMID: 11948787]

[06] M. Nooren and J. M. Thornton, *EMBO J.,* 22: 3486 (2003) [PMID: 12853464]

[07] K. Nadassy *et al.*, *Biochemistry,* 38: 1999 (1999) [PMID: 10026283]

[08] S. Jones *et al.*, *J Mol. Biol.,* 287: 877 (1999) [PMID: 10222198]

[09] N. M. Luscombe *et al.*, *Genome Biol.,* 1: 001.1 (2000) [PMID: 11104519]

[10] A. Sarai and H. Kono, *Annu. Rev. Biophys. Biomol. Struct.,* 34: 379 (2005) [PMID: 15869395]

[11] S. Biswas *et al., Proteins,* (2008) [PMID: 18704949]

[12] S. Jones *et al.*, *Nucleic Acids Res.,* 29: 943 (2001) [PMID: 10222198]

[13] M. Treger and E. Westhof, *J. Mol. Recogn.,* 14: 199 (2001) [PMID: 11500966]

[14] J. J. Ellis *et al., Proteins*, 66: 903 (2007) [PMID: 17186525]

[15] D. Endy, *Nature,* 438: 449 (2005) [PMID: 16306983]

[16] M. Guharoy and P. Chakrabarti, *Bioinformatics,* 23: 1909 (2007) [PMID: 17510165]

[17] G. Wang and R. L. Dunbrack, *Bioinformatics,* 19: 1589 (2003) [PMID: 12912846]

[18] H. M. Berman *et al.*, *Biophys. J.,* 63: 751 (1992) [PMID: 1384741]

[19] W. Kabsch and C. Sander, *Biopolymers,* 22: 2577 (1983) [PMID: 6667333]

[20] P. Chakrabarti and D. Pal, *Prog. Biophys. Mol. Biol.*, 76: 1 (2001) [PMID: 11389934]

Edited by P. Kangueane

Citation: Biswas *et al.*, Bioinformation 2(10): 422-427 (2008)

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Supplementary material

Formulae

$$(P_i)^{SSE} = (n_{i,sse,int}/N_{sse,int}) / (n_{i,sse,total}/N_{sse,total}) \rightarrow \quad (1)$$

where $n_{i,sse,int}$ and $N_{sse,int}$ are the counts of residue *i* and of all residues belonging to a particular secondary structure type in the interface, respectively; $n_{i,sse,total}$ and $N_{sse,total}$ are the corresponding counts in the entire tertiary structure.

Tables

PDB	Res (Å)[a]	Chain ID (length)[b] Amino acids	Nucleotides	Complete assembly[c]
1a9n	2.4	B(96)	Q(24)	1+
1c0a	2.4	A(585)	B(77)	1
1dfu	1.8	P(94)	M(19),N(19)	1
1di2	1.9	A(69)	C(10),D(10)	2
1f7u	2.2	A(607)	B(76)	1
1ffy	2.2	A(917)	T(75)	1
1fxl	1.8	A(167)	B(9)	1
1gtf	1.8	A(74)	W(55)	22
1hc8	2.8	A(76)	C(58)	2
1hq1	1.5	A(105)	B(49)	1
1i6u	2.6	A(130)	C(37)	2
1jid	1.8	A(128)	B(29)	1
1k8w	1.9	A(327)	B(22)	1
1knz	2.5	A(164)	W(5)	8
1m8x	2.2	A(349)	C(8)	2
1ooa	2.5	A(326)	C(29)	2
1pgl	2.8	2(185)	3(6)	2 * 60
1q2r	2.9	A(386)	E(44)	4
1qf6	2.9	A(642)	B(76)	1
1r9f	1.9	A(136)	B(21),C(21)	1
1sds	1.8	A(117)	D(15)	3
1si2	2.6	A(149)	B(9)	1
1u0b	2.3	B(461)	A(74)	1
1wpu	1.5	A(147)	C(7)	2
1xok	3.0	D(26)	A(30),B(9)	2
1yvp	2.2	A(538)	G(10)	2
1zbh	3.0	A(299)	E(20),F(20)	4
1zbi	1.9	A(142)	C(12)	2
1zh5	1.9	B(195)	C(9)	2
1zjw	2.5	A(553)	B(75)	1
2anr	1.9	A(178)	B(25)	1
2asb	1.5	A(251)	B(11)	1
2az0	2.6	A(73)	C(18),D(18)	2
2b3j	2.0	A(159)	E(16)	4
2bgg	2.2	A(427)	P(8),Q(8)	2
2bh2	2.2	A(433)	C(37)	2
2bu1	2.2	C(129)	S(19)	3

2db3	2.2	A(434)	E(10)	4
2dra	2.5	A(437)	B(34)	1
2dxi	2.2	A(468)	C(75)	2
2f8k	2.0	A(88)	B(16)	1
2fz2	2.9	C(189)	D(3)	3 * 60
2g4b	2.5	A(172)	B(7)	1
2gic	2.9	D(422)	R(45)	5 * 2
2gjw	2.9	A(313)	E(19),F(12)	4
2gxb	2.3	A(66)	E(7)	2
2hw8	2.1	A(228)	B(36)	1
2ipy	2.8	A(888)	C(30)	2
2q66	1.8	A(525)	X(5)	1
2uwm	2.3	A(258)	D(23)	2

Table 1: List of protein-RNA complexes. [a]Resolution of the Xray data. [b]Corresponds to the basic unit of the protein involved. [c] Complete assembly indicates how many times the basic unit is repeated to generate the biological unit according to NDB **[18]**. A '+' sign indicates that the assembly contains additional protein chains.

Feature	Protein-protein	Protein-DNA	Protein-RNA
Interface area (Å^2) – total	1906±759	2039±886	2103±1285
– due to protein	953	1014±440	1071±677
No. of SSSs	9.6	11.0	14.2
No. of SSSs per 1000Å^2 interface area	11.0	10.8	13.3
No. of helices	2.1	3.4	4.3
Helix length	4.8	5.6	4.6
No. of strands	2.8	2.6	3.6
Strand length	2.4	2.5	1.9
No. of NR segments	4.7	5.5	6.3
Length of NR segments	3.3	3.5	3.2

Table 2: Statistics on secondary structural segments in interfaces. For PP complexes both the components were taken together to get the average values **[16]**. These have been halved to facilitate comparison with the protein component of PD and PR complexes. Interface area for PP complexes is taken from **[5]**; the value "due to protein" corresponds to one component only.