BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis
Volume 8(1)

Accommodation of profound sequence differences at the interfaces of eubacterial RNA polymerase multi-protein assembly

Lakshmipuram Seshadri Swapna[1], Nambudiry Rekha[1] & Narayanaswamy Srinivasan[2]*

[1]Biobase Databases India Pvt Ltd, Langford Town, Bangalore 560025, India; 2Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560012, India; Narayanaswamy Srinivasan - Email: ns@mbu.iisc.ernet.in; Phone: +91-80-22932837; Fax: +91-80-23600535;*Corresponding author

Received December 11, 2011; Accepted December 17, 2011; Published January 06, 2012

Abstract:
Evolutionarily divergent proteins have been shown to change their interacting partners. RNA polymerase assembly is one of the rare cases which retain its component proteins in the course of evolution. This ubiquitous molecular assembly, involved in transcription, consists of four core subunits (alpha, beta, betaprime, and omega), which assemble to form the core enzyme. Remarkably, the orientation of the four subunits in the complex is conserved from prokaryotes to eukaryotes although their sequence similarity is low. We have studied how the sequence divergence of the core subunits of RNA polymerase is accommodated in the formation of the multi-molecular assembly, with special reference to eubacterial species. Analysis of domain composition and order of the core subunits in >85 eubacterial species indicates complete conservation. However, sequence analysis indicates that interface residues of alpha and omega subunits are more divergent than those of beta, betaprime, and sigma70 subunits. Although beta and betaprime are generally well-conserved, residues involved in interaction with divergent subunits are not conserved. Insertions/deletions are also observed near interacting regions even in case of the most conserved subunits, beta and betaprime. Homology modelling of three divergent RNA polymerase complexes, from *Helicobacter pylori*, *Mycoplasma pulmonis* and *Onion yellows phytoplasma*, indicates that insertions/deletions can be accommodated near the interface as they generally occur at the periphery. Evaluation of the modeled interfaces indicates that they are physico-chemically similar to that of the template interfaces in *Thermus thermophilus*, indicating that nature has evolved to retain the obligate complex in spite of substantial substitutions and insertions/deletions.

Keywords: RNA polymerase, eubacteria, homology modeling, obligate interactions, protein-protein interactions, sequence conservation, interface conservation

Background:
RNA polymerase (RNAP) is a ubiquitous molecular assembly which is indispensable for the process of transcription. This class of enzymes catalyzes the template-directed synthesis of RNA **[1]**. Specialized RNA polymerases have been fine tuned by evolution to synthesize different kinds of RNA. In eukaryotes, three nuclear RNA polymerases are present – RNAP I, RNAP II, RNAP III. RNAP I is required for the synthesis of pre-rRNA precursor of the three largest rRNAs **[2]**. RNAP II is involved in transcribing all mRNA along with non-coding RNAs **[3]**. RNAP III aids in synthesizing 5s rRNA, all tRNAs, and various short non-translated RNAs **[2]**. In eubacteria, a single molecule, evolutionarily closest to RNAP II, performs catalysis **[4]**. The archaeal RNAP resembles RNAP II holoenzyme **[5]**.

The multi-molecular assembly consists of four core subunits – alpha (I and II), beta, betaprime, and omega. These four subunits are common to RNA polymerase complexes of eubacteria, eukaryota and archaea. The sigma subunit aids in initiation of transcription in eubacteria (core enzyme + sigma = holo enzyme). The corresponding function is performed in eukaryotes and archaea by a combination of subunits. In addition to the above subunits, archaea and eukaryota contain additional subunits **[6]**. Viral RNA polymerases form exceptions to this structural organization by operating as single protein enzymes **[7]**. Remarkably, prokaryotic (*Thermus thermophilus* (PDB: 1IW7), *Thermus aquaticus* (PDB: 1HQM)) and eukaryotic (*Saccharomyces cerevisiae* (PDB: 2E2I)) holo enzyme structures exhibit high degree of structural similarity, although their sequence similarity is low **[8]**. Pairwise sequence identity for the core subunits between the RNA polymerases of these two organisms is in the range of 19-28%.

Study of interface conservation in proteins related at level of family/superfamily shows that as sequence divergence becomes extensive (as in protein domains related by superfamily), proteins tend to change their partners **[9]**. RNA polymerase subunits, although some of them show extensive divergence, seem to interact with the same partners. RNA polymerase system is thus a rare example. However, this is expected as the obligatory interaction between the various subunits is essential to successfully carry out transcription. The objective of this work is to investigate the structural accommodation of diverse sequences at the interface of RNA polymerase machinery. Further, we use homology modelling to understand how variations at the interface, such as substitutions, insertions and deletions in the sequences are accommodated in order to maintain the interface structure of the multi-molecular assembly.

Methodology:

Sequence analysis of homologous RNA polymerase subunits

Most of the subunits of RNA polymerase are multi-domain proteins. Two parameters were employed to estimate sequence divergence: variability in composition and order of sequence domains and percent sequence identity. Pfam **[10]** domain assignments of the different subunits of RNA polymerase complex of bacterial organisms with known genome sequences were extracted from the PRODOC **[11]** database. Greater than 100 sequences were obtained for alpha (121), beta (138) and betaprime (127) subunits each. 88 sequences were analyzed for omega subunit. Only 43 sequences were considered for sigma subunit, as only proteins assigned as sigma70, the general-purpose sigma factor were used in the analysis. The core subunits of RNA polymerases from all phyla were aligned using T-Coffee **[12],** to calculate sequence identities.

Identification of interface and surface residues

The interacting residues which are largely buried upon complexation are considered as interacting residues. These residues are identified on the basis of their residue surface accessibility (RSA) values: residues with RSA ≥10% in the free form (A) and RSA ≤7% in the bound form (AB) fulfill the criteria **[9]**. The surface accessibility of residues in each subunit was determined in two forms, free form and in complex with another subunit, using NACCESS **[13].** Surface residues were extracted by considering those residues in the free form of the subunit whose RSA ≥10% but which did not figure in the list of interface residues.

Estimating the extent of conservation of different residue types

The concept of Shannon Entropy (SE) was used to calculate the entropy of all interface residue positions. Shannon information analysis is a generalized mathematical tool to estimate variability **[14].** SE of a position (H) is calculated using the formula: $H= -\sum P(i) * \log_2 P(i)$ where H indicates SE of a position, P(i) indicates frequency of occurrence of 'i' in that position and 'i ' varies from 1 to 20 corresponding to 20 amino acid types. Literature survey **[14]** suggests the following cutoffs for H: 0.00 - 1.00 - conserved; 1.00 – 2.00 - semi-conserved. Analysis of different parameters suggests that an alignment containing ~100 sequences can be expected to be a good representative of a protein family for estimating variability using the Shannon measure. Around 130 sequences of alpha, beta and betaprime subunits of eubacterial species and 95 sequences of the omega subunit were used for the analysis. The eubacterial homologues of each core subunit were aligned using CLUSTALW **[15].** Shannon entropy of every position in the MSA was calculated using the formula specified above.

Modeling and evaluation of interfaces of divergent RNA polymerases

The pairwise alignments between template and model for each of the subunits were generated using FUGUE **[16],** PHYRE **[17]** and 3D-Coffee **[18].** The alignments were manually analyzed and refined where appropriate to arrive at a final alignment which was used as the basis for modeling. For cases of low sequence similarity between template and target sequences, the alignment was carefully analyzed and modified to ensure correspondence between observed and predicted secondary structures, in order to generate good models. The subunits were modeled individually using MODELLER **[19].** Small inserts were generated using the loop modeling module of MODELLER **[20].** The generated models were energy minimized using the Kollman United force field in SYBYL (Tripos Inc) to remove short contacts. After the individual subunits are modeled, we assembled the different subunits to form the RNA polymerase complex. Assuming that the topology of interaction between the various RNA polymerase subunits is preserved in all the eubacterial species (which seems probable considering that even the prokaryotic and eukaryotic RNA polymerase complexes retain the same interaction topology between their core subunits), the modeled subunits were assembled into a complex based on the layout of the *Thermus thermophilus* complex as template. The program SUPER (B.S. Neela, unpublished) was used to place the individual modeled subunits in the same orientation and location as they are present in the template. The entire complex was again energy minimized using SYBYL to remove short contacts.

The interface residues of every pairwise interface between the different subunits of the template RNA polymerase complex are extracted and its overall interface score calculated according to the formula: $Score = (\sum n_{ij} * R_{ij})/ Tot\_int$ where Score = Overall score for the interface, n_{ij} = number of interactions where amino acids i and j are in the environment of each other, R_{ij} = preference score for the case where amino acids i and j are in the environment of each other calculated from the reference non-redundant dataset **(Figure 1)**, Tot_int = total number of interacting residues in the pair-wise interface. The same process is repeated for all the interfaces of the RNA polymerase subunits from the modeled complexes. Those interfaces whose interface score falls within ±5% of the interface score of the template are considered to be similar.

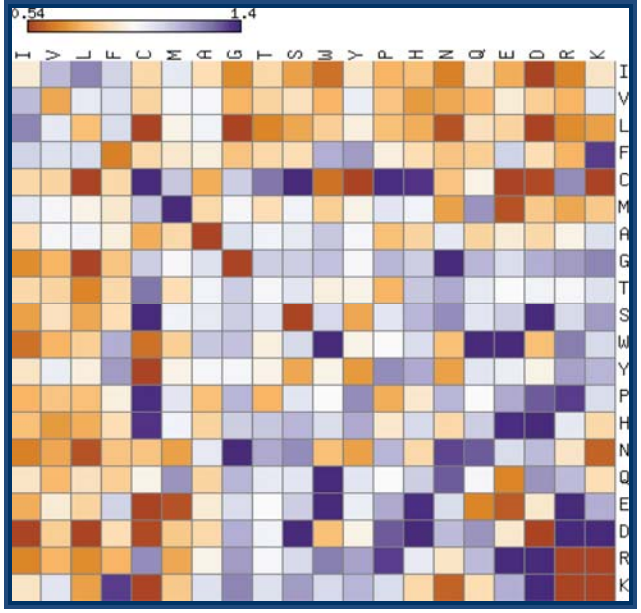

Figure 1: Log-odds matrix for determining preferred environment of the 20 amino acids. The reference matrix to indicate preferred environment of amino acids was generated as follows: The list of all interacting domains was culled out from SCOP 1.67 **[22],** and a non-redundant dataset comprising only one representative domain-domain entry for each SCOP family (the pair with the best resolution was chosen) was derived. The interacting residues in each complex were identified using RSA cutoffs as defined in Methods section. The residues in its environment were identified as those occurring within a Cβ-Cβ distance of ≤9Å. From this dataset, the preference for every amino acid to occur in the environment of each of the other 20 amino acids was calculated using the formula: $\text{Preference}_{ab}=\log_2 (P_{ab}/P_aP_b)$. The symbols 'a' and 'b' represent 2 amino acids in the environment of each other. P_{ab} represents the observed probability of occurrence of 'a' and 'b' in the environment of each other (calculated from the dataset). P_aP_b represents the expected probability of occurrence of 'a' and 'b' in the environment of each other. This log-odds score gives an idea of the preference of amino acid 'a' to be in the environment of amino acid 'b'. The preference score is calculated for each of the amino acid pairs (210 pairs) and represented in the form of a 20*20 matrix.

Discussion:

Extent of sequence divergence of core RNA polymerase subunits

The extent of sequence divergence in the four core subunits among members of all kingdoms was studied using two parameters: %sequence identity between homologous pairs and variation in domain composition and order.

The lowest percent sequence identities for homologous subunits across kingdoms are as follows: Alpha I (11%, *Nanoarchaeum equitans* (archaea) – *Synechochoccus sp* (eubacteria)), Alpha II (19%, *Halobacterium sp* (archaea) – *Homo sapiens* (eukaryota)), Beta (21%, *Bartonella henselae* (archaea) – *Plasmodium falciparum* (eukaryota)), Betaprime (20%, *Mycoplasma pulmonis* (eubacteria) – *Saccharomyces cerevisiae* (eukaryota)), Omega (8%, *Methanococcus jannaschi* (archaea) – *Candidatus Blochmannia floridanus* (eubacteria)). These values indicate the extensive sequence divergence of the subunits. However, when we consider RNA polymerase homologues only within the eubacterial species, the lowest percent sequence identities are as follows: Alpha (24%, *Helicobacter pylori* – *Mycoplasma pulmonis*), Beta (39%, *Helicobacter hepaticus* – *Mycoplasma pneumoniae*), Betaprime (38%, *Deinococcus radiodurans* – *Mycoplasma pulmonis*), and Omega (11%, *Candidatus Blochmannia floridanus* – *Treponema pallidum*). In eubacterial species, both alpha I and alpha II are identical. Within the eubacterial kingdom, RNA polymerase complex seems conserved, except for omega subunit, which shows extensive divergence.

Further, the various sequence domains of the core subunits were analyzed to determine whether any variability exists at the level of the individual Pfam sequence domains (addition or deletion of domains / difference in the order of domains), which generally serve as independent evolutionary and functional units. Analysis of the domain assignments reveals that domains and their order are completely conserved in all the subunits. Next, the interface residues for the various subunits, extracted from the RNA polymerase holoenzyme complex of *Thermus thermophilus* (PDB ID: 1IW7), were mapped on to the sequences to get an idea of the extent of participation of every sequence domain in interface formation. For each sequence domain, its variation in the members of the kingdom was analyzed in terms of its length in an attempt to correlate the variation in length (indicating insertions/deletions) with the extent/nature of the sequence domain's participation in interaction with other subunits **(Table 1, see supplementary material).**

The analysis indicates that no major insertions/deletions are present in the alpha subunit or omega subunit. In the case of sigma70 subunit, the Pfam-assigned domains for the sigma70 subunit of *Thermus thermophilus* are completely conserved. However, many other sigma70 members contain two more Pfam domains (which correspond mainly to unassigned region in *Thermus thermophilus*). Both do not seem to be involved in interface formation. The Pfam domain corresponding to region 1.1 of sigma70 factor (PF03979), found at N-terminus, modulates DNA binding to region 2 and 4 of the same subunit when RNA polymerase is not bound to the sigma70 subunit. Another Pfam domain (PF04546), found in the primary vegetative sigma factor, is a non-essential region. Huge insertions/deletions are present in beta and betaprime subunits, the best conserved of the RNA polymerase subunits, even in domains involved in interaction with other subunits **(Table 1, italicized see supplementary material)**. This analysis indicates that insertions and deletions are common in several domains involved in interaction even in an evolutionarily well conserved multi-molecular assembly like RNA polymerase, corroborating other studies **[21].**

Extent of conservation of interface residues of RNA polymerase subunits

Based on the MSA of each of the four core enzymes, Shannon Entropy (SE) was calculated for three categories of residues: all (overall), interface and surface residues. Initially, % conserved residues (see Methods) in a particular residue class (interface / surface / overall) was calculated using SE method **(Table 2, see supplementary material)**. As expected, a higher proportion of interface residues are conserved when compared to surface residues or the whole protein. Also, surface residues are slightly less conserved when compared to the whole protein. The order of subunits in terms of the percent of interface residues conserved is Beta >~ Betaprime > Alpha II > Alpha I > Omega **(Table 2, see supplementary material).** However, when we consider the Interface/Surface ratio, the trends completely reverse. Omega subunit, which has the lowest Overall and Surface conservation, has the highest Interface/Surface ratio value. This is followed by Alpha I and Alpha II, the subunits with intermediate % Overall conservation, finally followed by Betaprime, Beta, and Sigma70, the most conserved subunits. This trend is in keeping with the fact that beta and betaprime should have more conserved residues on the surface (apart from those involved in interface formation), as the subunits are critical for catalysis, DNA binding, RNA binding and substrate binding **[6]**. In contrast, the main function of alpha and omega subunits seems to be interaction with other core subunits to form RNA polymerase assembly. This facet seems to be indicated by the high Interface/Surface ratio of these more divergent subunits. In summary, a larger proportion of interface residues of the more divergent RNA polymerase subunits (alpha, omega) are conserved than the rest of the surface in comparison to that of the less divergent (beta, betaprime, sigma70) RNA polymerase subunits.

Next, we analyzed the conservation of specific pair-wise interfaces of RNA polymerase subunits using SE measure **(Table 3, see supplementary material)**. The trend seen from the Interface/Surface values for the various pair-wise interfaces corroborates the pattern seen for Overall Interface residues. The more divergent subunits show better capability to distinguish their interfaces rather than the more conserved subunits. Another interesting observation from the analysis is that although beta and betaprime are generally well-conserved, the residues involved in interaction with the divergent subunits are not well conserved **(Table 3, see supplementary material)**. In light of evidence that the interface structure is maintained in evolution, this indicates that these interacting positions, although not conserved, probably co-evolve to maintain the interface structure.

Modeling and evaluation of interfaces of three divergent eubacterial RNA polymerases

From the preliminary analysis of the multiple sequence alignment, it is evident that there exist cases of large insertions and deletions near to the interface in even the most conserved RNA polymerase subunits (beta and betaprime). Therefore, we explored, using modeling, how insertions/deletions at the interface **(Table 1, see supplementary material)** and substitutions at the interface, which account for 30% (beta, betaprime) - 85% (omega) changes in interface residues **(Table 2, see supplementary material)**, are accommodated in the formation of the complex and evaluate the similarity of the modeled interface with that of the template's interface (using log-odds matrix, **Figure 1**). The species chosen for modeling are: *Helicobacter pylori*, *Onion yellow phytoplasma* and *Mycoplasma pulmonis*. These were chosen because they contained the poorest %sequence identity (with respect to majority of the RNA polymerase subunits) and contained some interesting insertions/deletions near the interface. The RNA polymerase holoenzyme complex of *Thermus thermophilus* (PDB: 1IW7) solved at 2.6 Å was used as the template structure. The sequence identities between template and target sequences for all subunits are listed in **(Table 4, see supplementary material)**. The locations of the insertions / deletions occurring in the sequences of the target subunits are mapped onto the crystal structure of the template in **Figure 2**. We observe that some insertions and deletions (indicated using arrows in **Figure 2**) occur near subunit-subunit interfaces. However, all the insertions occur at the periphery of the assembly, indicating that they can be accommodated in spite of the large size.

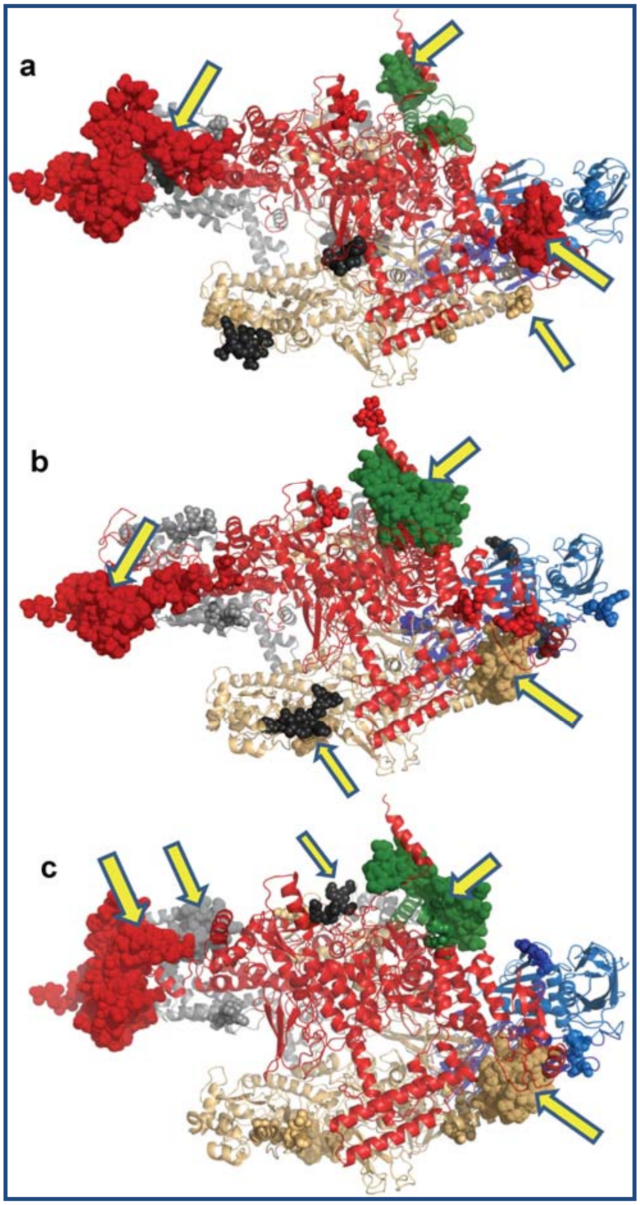

Figure 2: Insertions and deletions in the subunits of **a)** *Helicobacter pylori* **b)** *Onion yellows phytoplasma* and **c)** *Mycoplasma pulmonis* mapped onto the crystal structure of the

macro-molecular assembly of RNA polymerase from *Thermus thermophilus* (PDB:1IW7). The subunits are colored as follows: alpha I (dark blue), alpha II (marine blue), beta (yellowish orange), betaprime (red), omega (green), sigma70 (grey). Regions containing small insertions in the subunits are indicated as spheres colored according to the respective subunit, regions containing large insertions are indicated as black spheres and deletions are shown as large dotted spheres. Insertions / deletions which occur close to subunit-subunit interfaces are indicated using yellow arrows.

The holoenzyme generated for the three organisms (see organisms) was used for further analysis of the interfaces formed between the different subunits in the modeled RNA polymerase complex. The residues participating in interface formation were extracted for both template and modeled complexes using accessibility information (see Methods). All the residues from the partner chain whose Cβ atom (Cα in case the residue is glycine) is ≤9 Å from the Cα atom (Cα for Gly) of the interface residue is considered to be in the environment of the latter. Cβ-Cβ distance is chosen instead of atom-atom distances as the latter parameter may not be reliable in case of modeled structures due to inaccuracies in side-chain positioning.

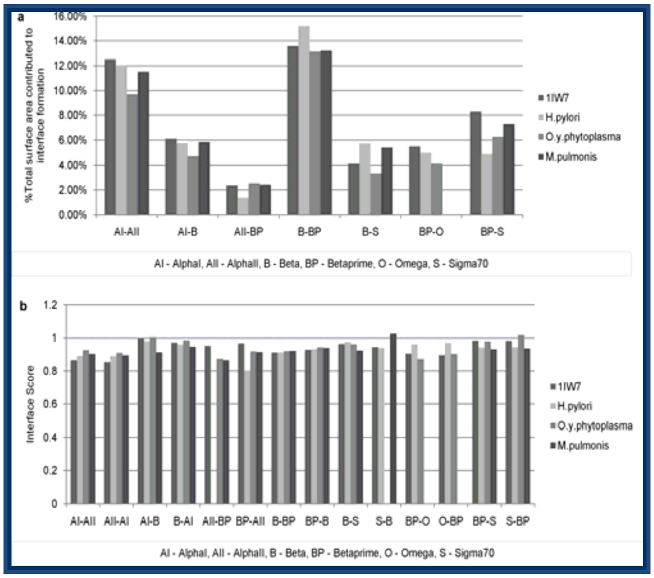

Figure 3: Assessing the similarity between interfaces of the template and modeled holo-enzyme RNA polymerase structures. Similarity is assessed using **(a)** %Total surface area involved in interface formation **(b)** Interface score for all pair-wise interfaces in template (1IW7) and modeled structures.

Figure 3a shows the distribution of the %surface contributed to interface formation between various subunits of the RNA polymerase complex. In the case of the betaprime-omega interface, although similar surface area equivalent to one observed in template structure is involved in interface formation for *Helicobacter pylori* and *Onion yellows phytoplasma*, this interface is absent in *Mycoplasma pulmonis* because the omega subunit is absent in the latter. For the rest of the pairwise interfaces, the %surface area contributed to interface formation is almost similar to the area buried in the template structure, indicating that similar sized interfaces have been modeled. Slight variations can occur due to the small insertions / deletions at interface regions. We also note that all large insertions and deletions could be accommodated since they occur on the periphery and would be amenable to modeling if a suitable template was available for those regions. Next, we compare if the modeled interfaces are physico-chemically similar to the corresponding interface in the template structure. **Figure 3b** depicts the comparison of overall interface scores for the template and modeled RNA polymerase interfaces. Apart from alpha II – betaprime interface (of *Helicobacter pylori*) and sigma70 – beta interface of (*Onion yellows phytoplasma*) , we see that the interface scores of all other pairwise interfaces are similar to that of the corresponding values in the template in most of the cases, indicating that they are able to accommodate substitutions at the interface. The smaller insertions and deletions were modeled and do appear to be accommodated while maintaining the physico-chemical complementarity of the interface.

Conclusion:

The structure of RNA polymerase assembly is retained during the course of evolution. Although some of the core subunits of RNA polymerase complex show high sequence divergence, their interacting partners are retained. Furthermore, the orientation of the interacting partners is also conserved. This feature contrasts with the general behavior of homologous proteins, which change their interacting partners during extensive divergence. Analysis of domain composition and order of the core subunits of the RNA polymerase assembly in >85 eubacterial species indicates complete conservation. However, conservation analysis of the various core subunits indicates that the interface residues are more divergent for alpha and omega subunits. Although beta and betaprime are generally well-conserved, the residues involved in interaction with the divergent subunits (i.e. alpha, omega) are not conserved. Insertions/deletions are also observed near the interacting surfaces even in case of the most conserved subunits (beta and betaprime). Using homology modelling of three divergent RNA polymerase complexes, *Helicobacter pylori*, *Mycoplasma pulmonis* and *Onion yellows phytoplasma*, we observe that insertions/deletions can be accommodated near the interface as they generally occur at the periphery. Using a generalized matrix capturing preferences of interface environment, we find that the modeled interfaces are physico-chemically similar to that of the template interfaces in *Thermus thermophilus,* indicating that nature accommodates substantial substitutions and indels at and near the interface in order to retain the structure of the obligate complex, which is indispensable for the process of transcription.

Acknowledgements:

L.S.S is supported by the Mathematical Biology programme, Department of Science and Technology grant. This research is supported by the Mathematical Biology programme, Department of Science and Technology, Department of Biotechnology and Microsoft Corporation.

References:

[1] Cramer P & Arnold E, *Cur Opin Struct Biol.* 2009 **19**: 680 [PMID: 19910185]

[2] Werner M *et al*. *Cur Opin Struct Biol*. 2009 **19**: 740 [PMID: 19896367]

[3] Sydow JF & Cramer P, *Cur Opin Struct Biol.* 2009 **19**: 732 [PMID: 19914059]

[4] Vassylyev DG, *Cur Opin Struct Biol*. 2009 **19**: 740 [PMID: 19896365]

[5] Hirata A & Murakami KS, *Cur Opin Struct Biol*. 2009 **19**: 724 [PMID: 19880312]

[6] Cramer P, *Cur Opin Struct Biol*. 2002 **12**: 89 [PMID: 11839495]

[7] Steitz TA, *Cur Opin Struct Biol*. 2009 **19**: 683 [PMID: 19811903]

[8] Ebright RH, *J Mol Biol*. 2000 304: 687 [PMID: 11124018]

[9] Rekha N *et al*. *Proteins*. 2005 **58**: 339 [PMID: 15562516]

[10] Bateman A *et al*. *Nucleic Acids Res.* 2002 **30**: 276 [PMID: 11752314]

[11] Krishnadev O *et al*. *Nucleic Acids Res*. 2005 **33**: W126 [PMID: 15980440]

[12] Notredame C *et al*. *J Mol Biol*. 2000 **302**: 205 [PMID: 10964570]

[13] Lee B & Richards FM, *J Mol Biol.* 1971 **55**: 379 [PMID: 5551392]

[14] Stewart JJ *et al*. *Mol Immunol*. 1997 **34**: 1067 [PMID: 9519765]

[15] Thompson JD *et al*. *Nucleic Acids Res*. 1994 **22**: 4673 [PMID: 7984417]

[16] Shi J *et al*. *J Mol Biol*. 2001 **310**: 243 [PMID: 11419950]

[17] Kelley LA *et al*. *J Mol Biol*. 2000 **299**: 499 [PMID: 10860755]

[18] O'Sullivan O *et al*. *J Mol Biol*. 2004 **340**: 385 [PMID: 15201059]

[19] Sali A & Blundell TL, J Mol Biol. 1993 **234**: 779 [PMID: 8254673]

[20] Fiser A *et al*. *Protein Sci*. 2000 **9**: 1753 [PMID: 11045621]

[21] Lane WJ & Darst SA, **J Mol Bio**l. 2010 **395**: 671 [PMID: 19895820]

[22] Murzin G *et al*. *J Mol Biol*. 1995 **247**: 536 [PMID: 7723011]

Edited by P Kangueane

Citation: Swapna *et al.* Bioinformation 8(1): 006-012 (2012)

Supplementry material:

Table 1: Analysis of variation in length of Pfam domains constituting core RNA polymerase subunits

Core Subunit & Pfam domains	%Sequence domain involved in interface formation	Interacting partners	Minimum and maximum length	Mean length (± s.d)
Alpha				
PF01193	12.14	Alpha, Beta	204 – 229	211.48 ± 5.17
PF01000	6.8	Beta, Betaprime	113 – 134	117.98 ± 4.91
Beta				
PF04563	0	-	351 – 599	458.57 ± 49.94
PF04561	0	-	13 – 456	233.49 ± 89.63
PF04565	5.5	Betaprime	72 – 72	72 ± 0
PF00562	*10.4*	*Betaprime, Alpha I, Sigma70*	*388 – 618*	*478.79 ± 76.14*
PF04560	40.25	Betaprime, Sigma70	74 – 76	75.95 ± 0.28
Betaprime				
PF04997	*6*	*Sigma70, Beta*	*327 – 634*	*349.13 ± 52.3*
PF00623	21.23	Beta, Omega	120 – 151	142.24 ± 2.68
PF04983	*9.6*	*Beta, Alpha II*	*140 – 203*	*164.18 ± 15.99*
PF05000	6	Beta	80 – 104	87.32 ± 6.77
PF04998	*5*	*Beta, Omega*	*364 – 701*	*484.96 ± 103.47*
Omega				
PF01192	47.3	Betaprime	43 – 65	54.13 ± 4.33

#Domains involved in interaction with other subunits showing larger variation in length are highlighted in italics

Table 2: Shannon Entropy-based conservation of RNA polymerase subunit residues

RNA polymerase subunit family	OverallSE-Cons (%)	Surface SE-Cons (%)	Interface SE-Cons (%)	Interface SE-Cons
				Surface SE-Cons
Alpha I	21.87	15.65	52.17	3.33
Alpha II	21.87	18.35	64.29	3.50
Beta	34.29	35.15	68.67	1.95
Betaprime	36.41	31.53	68.07	2.15
Omega	6.38	3.51	13.33	3.79

Table 3: Analysis of specific interfaces of RNA polymerase subunits using Shannon Entropy

Interface	Overall	Surface	Interface	Interface SE-Cons
	SE-Cons (%)	SE-Cons (%)	SE-Cons (%)	Surface SE-Cons
Alpha I – Beta	21.87	15.65	57.14	3.65
Alpha I – Alpha II	21.87	15.65	44.44	2.83
Alpha II – Alpha I	21.87	18.35	50.00	2.72
Alpha II – Beta	21.87	18.35	100.00	5.45
Alpha II – Betaprime	21.87	18.35	100.00	5.45
Beta – Alpha I	34.29	35.15	60.00	1.74
Beta – Sigma70	34.29	35.15	100.00	2.91
Beta – Betaprime	34.29	35.15	67.65	1.92
Betaprime – Omega	36.41	31.53	22.22	0.70
Betaprime – Sigma70	36.41	31.53	87.50	2.77
Betaprime – Alpha II	36.41	31.53	0.00	0
Betaprime – Beta	36.41	31.53	73.81	2.34
Omega – Betaprime	6.38	3.51	13.33	3.79
Sigma70 – Beta	24.84	41.95	100.00	2.38
Sigma70 – Betaprime	24.84	41.95	47.06	1.12

Table 4: Sequence identity between Template (*Thermus thermophilus*) and Target subunits

Organism	Alpha I	Beta	Betaprime	Omega	Sigma70
Helicobacter pylori	21.45	36.51	34.73	7.37	23.29
Mycoplasma pulmonis	30.24	41.92	31.72	- #	38.16
Onion yellows phytoplasma	40.34	36.89	34.36	5.35	33.19

#In *Mycoplasma pulmonis,* no protein corresponding to omega subunit has been identified