academicJournals

Vol. 5(2), pp. 23-43, July 2013
DOI: 10.5897/JCBBR10.011
ISSN 2141-2227© 2013 Academic Journals
http://www.academicjournals.org/JCBBR

Journal of Computational Biology and Bioinformatics Research

Full Length Research Paper

Does a protein coevolve with its multiple interacting partners? A case study

Subinoy Biswas and Sudip Kundu*

Department of Biophysics, Molecular Biology and Bioinformatics, University of Calcutta, Kolkata 700 009, India.

Accepted 7 February, 2013

Protein-protein interactions are playing a fundamental role in different cellular activities. Although the coevolution of interacting protein pairs has been established by several groups, whether a protein having multiple interacting partners coevolves with all of its interacting partners or not have not been studied, so far. Here, the coevolution of proliferating cell nuclear antigen (PCNA) with their multiple interacting protein partners was studied. The 'mirror tree' method was used to predict the signature of coevolution of the interacting pairs. The results show that PCNA, which interacts with a larger number of proteins, does not coevolve with each of its partners. Rather, the degree of coevolution varies in a statistically significant wider range. The nature of coevolution of these interacting pairs in two different lineages (archaea and eukarya) has been further investigated separately. Results show that the coordinated evolutions of some of the interacting pairs are different for two different lineages. The possible reasons (percentage of disorder region of partner proteins, synonymous to non-synonymous ratio, cascade interactions, etc.) of the variations have also been discussed.

Key words: Proliferating cell nuclear antigen (PCNA), coevolution, protein-protein interaction.

INTRODUCTION

Proteins rarely act alone. A large number of proteins interact with other proteins to carry out their respective biological functions (Pereira et al., 2006; Grigoriev, 2003) and several studies have focused on such protein-protein interactions with emphasis on their different structural and functional properties including preference of residues at the interface, combinatorial effect of interactions, the emerging properties of protein-protein interaction networks (PPIs), etc (Bork et al., 2004; Argos 1988; Janin et al., 1988; Jones and Thornton 1996; Hoskins et al., 2006; Pal et al., 2006). In addition to the structural and functional perspective, numerous studies have attempted to identify the trends in the evolution of such interacting protein partners (Altschuh, et al., 1987; Moyle et al., 1994; Pazos et al., 1997; Goh et al., 2000). These studies have shown that in the case of systems containing two different interacting proteins, change in one interacting partner often imparts a direct influence on evolution, often through a compensatory change, in the other partner to maintain the structural and functional integrity of the complex. Moreover, even in cases where the interactions among the different domains of the same protein is known to be important for its biological functions, these interacting domains have been generally observed to be coevolved that is a heritable change in one of the interacting domain has been found to exerts a selective pressure for a corresponding change in other interacting domain(s).

However, most of our knowledge of the nature of coevolution (the term 'coevolution' has been used to refer to the similarity of evolutionary histories, which can be quantified through the similarity of the corresponding phylogenetic trees of proteins) of proteins are based on studies on systems containing paired interacting protein

*Corresponding author. E-mail: skbmbg@caluniv.ac.in. Tel: 91-33-23508386. Fax: 91-33-23519755.

partners. However, as a large number of cellular proteins are known to interact with multiple interacting partners (at least some of which may in turn interact with one or more interacting partners), it is significant to investigate whether the observed trends for coevolution of paired interacting partners remain valid for evolution of proteins which are involved in complex protein interaction networks that are common in nature. To address the largely unexplored problem of evolution of proteins in context of such complex interaction network architectures, we have used the evolutionary analysis of the Proliferating Cell Nuclear Antigen (PCNA) and its interacting partners, to assess the extent of structural and functional constraints that may be imposed on the evolution of a protein due to its interaction with different interacting partners.

The PCNA, is a member of the so-called DNA sliding clamp family which has a remarkable ability to interact with multiple proteins (Giovanni and Ulrich, 2003). The interacting partners of PCNA interact with PCNA through different but specific interacting sites. The sites are mainly the inter-domain connecting loop of PCNA ring like structure, N-terminal region comprising inner α - helices and the C terminal tail of PCNA (Jonsson and Hubscher, 1997; Warbrick, 2000). Although PCNA is known to interact with numerous partners, only ten of its interacting protein partners (Replication factor C3(RFC3), DNA Polymerase delta(pold), DNA Ligase 1(Ligase 1), DNA Topoisomerase 1(Topo 1), DNA Topoisomerase 2(Topo 2), Flap endonuclease 1(Fen 1), XPG endonuclease(XPG), WRN helicase(WRN), MLH 1(MLH 1), Uracil-DNA glycosylase(Uracil)) for which comprehensive literature based evidence for physical interactions of these proteins with PCNA as well as corresponding protein sequences from various taxa available, were selected for the present study.

We observed that PCNA does not have similar correlated evolution with all of its ten interacting partners. Rather, the values of correlation coefficients indicate varying degrees of correlated evolution of PCNA with its interacting partners. This lead to notion that a protein having multiple number of interacting protein partners may not coevolved with all of its partners. We have further studied the correlated evolution in two different lineages: eukarya and archaea separately. Significant differences have been observed in two lineages for some of the interacting partners. We have also searched for the possible underlying reasons for different values of correlation coefficients. When varying number of interacting partners do not throw any light, the degree of disorder of the interacting protein partners exhibit some clue for it. Here, we have explored the possibility of any specific signature that may correlate the nature of coadaptation with percentage of disorder region of the interacting partners. We have further extended our search by measuring nonsynonymous (dn) to synonymous (ds) ratio to understand whether these values can provide any rationale for the observed variations in the degrees of co evolutionary pressures.

MATERIALS AND METHODS

Data collection

Protein sequences of PCNA and its ten interacting partners (proteins) from nine eukaryotic and nine archaeal species were collected from the NCBI database (http://www.ncbi.nlm.nih.gov) and are listed in Supplementary Table S1. All the proteins were collected by protein name query in the NCBI database. Whenever a desired protein was not found by simple name query, protein blast (BLASTP) (Altschul et al., 1990) in NCBI followed by manual curations was performed to incorporate such sequences in our study (Supplementary Table S1 1a, 1b). All these sequences form the dataset 1. The dataset 2, which is a subset of dataset 1, includes only those proteins from the dataset 1 which are properly annotated that is neither hypothetical nor putative (putative and others are marked) (Supplementary Table 1a and 1b).

When we have studied the coevolution of PCNA with its ten interacting partners (proteins) in two different lineages separately, a comparatively larger set of sequences was used. These are listed in Supplementary Table S2 and denoted as dataset 3. All the sequences of dataset 3 were properly annotated. It includes all the sequences present in dataset 2 and also some sequences from NCBI and Orthodb (http://cegg.unige.ch/orthodb2). While protein sequences of four interacting proteins (MLH1, Uracil, XPG and WRN) of eukaryal species were taken from database of orthologous groups (http://cegg.unige.ch/orthodb2), the rest of the protein sequence of interacting partners were collected from NCBI. Contrary to the Table S1, all the sequences of the interacting proteins were not from the same set of species. However, the coevolution of any interacting pairs was studied using the sequences taken from the same set of species.

To calculate the dn/ds ratio, we collected the respective DNA sequences of proteins (listed in dataset 3) from NCBI.

Calculation of correlation coefficient (r) as an indicator of coevolution

To measure the correlated evolution of interacting partners, the most widely used method (Goh et al., 2000; Pazos and Valencia, 2001; Goh and Cohen, 2002; Ramani and Marcotte, 2003; Kim et al., 2004; Tan et al., 2004; Pazos et al., 2005; Sato et al., 2005; Mintseris and Weng, 2005; Pazos and Valencia, 2008; Pazos et al., 2008) "entire-sequence" approach of “mirror tree” comparison was used. In this method, pair wise distance matrices derived from the alignment of entire amino acid sequences were compared, their correlation coefficient values were calculated and the detections of statistically significant correlations were used to infer correlated evolutions.

Sequences of the two interacting proteins have been taken from the same set of species. CLUSTALW (Higgins et al., 1994) was used to align the sequences. The distance matrices were calculated using PROTDIST of PHYLIP (Felsenstein, 2002) package with Jones-Taylor-Thornton matrix. The linear correlation coefficient of these two distance matrices was calculated using the expression (Press et al., 1992).

$$r = \frac{\sum_{i=1}^{n}(R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_{i=1}^{n}(R_i - \overline{R})^2}\sqrt{\sum_{i=1}^{n}(S_i - \overline{S})^2}}$$

Where n is the number of elements of the matrices, that is, $(N^2-N)/2$, N is the number of sequences in the multiple sequence alignments, Ri are the elements of the first matrix (the distances among all the proteins in the first multiple sequence alignment), Si is the corresponding value for the second matrix and $\overline{R}$ and $\overline{S}$ are the respective average of Ri and Si, respectively. It should be mentioned that this r-value is an indicator of coevolution. The higher the r-value (positive) represents the more coordinated evolution.

A bootstrap analysis is used to estimate the statistical significance of the computed correlation coefficient values (r). For this, we generated 1000 sets containing n pair-wise distances randomly drawn (with replacement) from the n pair-wise distances in the original set and calculated 1000 values r_{rand}. Z score was calculated using the expression:

$$Z = \frac{r - \bar{r}_{rand}}{\sigma_{rand}}$$

Where, σ is the standard deviation of r_{rand} and $\overline{r}_{rand}$ is the mean (effectively zero for truly random data). The p-value is then obtained from $p = \text{erfc}(|z|)/\sqrt{2}$, where erfc is the complement error function.

Further, we also used a two-tailed test to predict whether any two calculated r-values are statistically significantly different or not (Spiegel, 1972).

An in-house PERL script is used to calculate the r-values and the corresponding p values.

Phylogenetic tree building

For a given set of orthologous sequences, we first generated the multiple alignment using CLUSTALW (version 1.83), a progressive alignment method. For generating the bootstrapped tree, we generated the multiple copies using seqboot, and distance matrices were calculated using PROTDIST with Jones-taylor-Thornton matrix. The phylogenetic trees were constructed for multiple data sets using NEIGHBOR, a neighbor-joining method. The final tree for each of the proteins was generated using CONSENSE program. We used Phylip package version 3.6.

Protein disorder calculations

Evidence is rapidly accumulating that many protein regions and even entire proteins lack stable tertiary and/or secondary structure in solution yet possess crucial biological functions. These naturally flexible proteins regions are known by different names. We refer to these flexible regions as protein disorder region in this article. Protein disorder region provides essential biological functions because dynamic conformation allows proteins to interact with multiple targets (Dunker et al., 2002). Disordered regions are comprised of a category of amino acids distinct from that of ordered protein structures (Garner E, Cannon P,Genome Inform Ser Workshop Genome Inform 1998). We used a well established web server Poodle-S (http://mbs.cbrc.jp/poodle/poodle-s.html) (Kana Shimizu et al., 2007) to calculate protein disorder region and from that we calculated the percentage of disorder region of two eukaryotic organisms and three archaeal organisms for all 10 PCNA interacting partners, which we considered in our study.

Dn/Ds calculations

Estimation of nonsynonymous and synonymous substitution rates is widely used to understand the dynamics of molecular sequence evolution (Gillespie, 1991; Ohta, 1995). We used yn00 program of Paml3.14 package for dn/ds calculation following Yang and Nielsen (2000) method of estimation. We used the maximum likelihood method for pairwise sequence comparison. When nonsynonymous (dn) to synonymous (ds) ratio(ω) is <1, =1, >1, it is a negative selection, neutral and positive selection respectively. Coevolved interacting partners tend to show (ω) <1 selection pressure on them due to evolutionary conservation. Partners showing (ω) >1 in was the case of proteins which is not under evolutionary constrained and positive selection acting on those proteins.

RESULTS AND DISCUSSION

We calculated the Pearson correlation coefficient (r) values of PCNA and each of its ten interacting protein partners. The accession numbers of those protein sequences (dataset 1) are given in Supplementary Table S1, a and b). The r-values are listed in Table 1. We also calculated the statistical significances of these r-values. The result shows that all the r-values except the values marked as # have p values of less than 10^{-5}. The results show that seven among ten interacting partners of PCNA, namely Ligase 1, Pold, Fen 1, Topo 1, Topo 2, MLH 1 and Uracil, had high correlation coefficient values ($r > 0.6$). On the other hand, WRN had comparably smaller r-values, whereas the other two interacting partners of PCNA, namely XPG endonuclease and RFC3 show very low and negative correlations (almost no correlation), respectively with PCNA.

To study how correlated evolution act on different interacting partners of PCNA, a well-established entire sequence based correlation coefficient value approach was employed. As all the proteins included in our study are known interacting partners (Giovanni and Ulrich 2003), we can excluded the possibility of false positive results that may arise due to chance. A well established method like 'mirror tree' approach is used to study the pattern of evolution of PCNA with their interacting partners. It should be mentioned that the aim of this paper was not to find any new interacting partner, but to understand the evolutionary relationships of the interacting partners with PCNA.

It is expected that the interacting proteins should coevolved (Atwell et al., 1997; Jespers et al., 1999; Moyle et al., 1994; Pazos et al., 1997). The high value of correlation coefficient (r) of two interacting proteins is an indicator of this correlated evolution of the partners (Goh et al., 2000). The interaction of PCNA with each of the ten proteins, included in our study, is experimentally verified (Giovanni and Ulrich, 2003).

So, we expect high positive r-values for each of the ten partners. However, we observed a wide range of r-values starting from very low negative (nearly zero) to high positive r-values. The statistical significances of the wide variation of the calculated r-values are given in Supplementary Table S3 to S8). This indicates that there is different order of constraints acting on PCNA and its different interacting partners. It should be mentioned that

Table 1. Correlation coefficient values of PCNA and its ten different interacting partners.

Interacting partner	*r-value*	
	Archaea + Eukarya combined	
	Including hypothetical	Without hypothetical
Ligase 1	0.759	0.752(17)
Pold*	0.711	0.897(12)
Topo 1	0.806	0.749(17)
Topo 2	0.792	0.848(17)
Fen1	0.837	0.842(17)
RFC3	-0.056	0.202(16)
MLH 1*	0.637	0.690(12)
XPG	-0.089	-0.030(17)
WRN	0.394	0.520(11)
Uracil*	0.630	0.64(17)

Ligase 1, DNA Ligase 1; Pold, DNA Polymerase delta; Topo 1, DNA Topoisomerase 1; Topo 2, DNA Topoisomeras 2; Fen 1, Flap endonuclease 1; RFC3 , Replication factor C3; MLH 1, MLH 1(mismatch repair protein); XPG, XPG endonuclease; WRN, WRN helicase; Uracil, Uracil DNA Glycosylase. No Creanarchaea organisms were used in studying the coevolution for the interacting partners marked with *. The numbers within parenthesis represent the sample size. All the *r*-values are statistically significant with $p \leq 10^{-5}$.

some of the sequences (taken from 18 different species) used in the above study (dataset 1) are hypothetical, putative, etc, that is there is no experimental evidence of their functional annotation. Therefore as a next step, we calculated the r-values using only those sequences that are neither hypothetical nor putative (dataset 2).

In almost all the cases, except two (Topo 1 and Ligase 1), the r-values showed a clear increase (Table 1). Most significantly, the r-value of RFC3 becomes positive (0.202) as is evident from Table 1. However, it is still significantly low to conclude strong coordinated evolution of PCNA and RFC3. Interestingly, XPG still shows a very weak negative correlation (almost no correlation) with PCNA. Here, we also observed statistically significant differences in the r-values (Supplementary Table S4). It is also observed that when we use hypothetical or putative orthologs, we obtained comparably lower r-values. This is also expected because the hypothetical or putative orthologs have larger variations in their sequences.

It is evident from Table 1 that there is a wide variation in the r-values and the variations are also statistically significant (Supplementary Table S3 and S4). While seven interacting partners of PCNA exhibit different orders of constrains to maintain their coevolution with PCNA; three partners, namely RFC3, WRN and XPG do not show any coordinated evolution with PCNA. The results can be explained by the following arguments. The protein, PCNA has several interacting partners. The interacting partners may impose different evolutionary pressures depending on the necessity of structural and functional integrity of each of the interacting complexes. Thus, the result supports our hypothesis that a protein having multiple interacting partners may not coevolved with all of its partners, even the degrees of evolutionary pressures (constrained imposed to any change) may vary in a wide range.

So how do the coevolution of PCNA and their interacting partner proteins along two different lineages (archaea and eukarya separately) follow? Phylogenetic analysis of all available archaeal PCNA homologues suggests that creanarchaeal homologs are divided into two groups while other archeal PCNA have single PCNA (Toshie et al., 2000). So, to keep homogenity of PCNA homologues in archaeal set, we exclude any crenarchaeal sequence that was previously considered in combined set. WRN homologues and Uracil homologues were not functionally annotated in most of the archaeal organisms that we considered earlier, hence not used in independent archaeal study.

The r-values (using dataset 3) (Supplementary Table S2 2a and 2b) obtained are listed in Table 2. We observed differences in r-values between archaea and eukarya in most of the cases. The statistical significances of the differences in r-values between archaea and eukarya lineages are given in Supplementary Table S5.

While the r-value obtained from eukaryal PCNA and polymerase delta is very high (0.897), the archaeal counterpart had lower r-value (0.693). The r-values show statistically significant difference ($p < 0.01$). The smaller value of r in the case of archaea and its significant differences with eukaryal r-value clearly indicate that the archaeal polymerase delta and PCNA evolved in a less coordinated manner than their eukaryal counterparts. The protein Fen1 also had significantly higher r-value in eukarya than that of archaeal counterparts ($p<0.01$). On

Table 2. Correlation coefficient values of PCNA and its ten different interacting partners.

Interacting partner	*r-value*	
	Eukarya	**Archaea**
Ligase 1	0.816(10)	0.796(12)
Pold	0.897(10)	0.693(12)
Topo 1	0.464(16)	0.889(12)
Topo 2	0.405(17)	0.573(12)
Fen 1	0.786(12)	0.563(12)
RFC3	0.235(10)	0.780(12)
MLH 1	0.835(17)	0.806(6)
XPG	0.368(15)	0.716(7)
WRN	0.569(16)	NS
Uracil	0.468(16)	NS

The numbers within parenthesis represent the sample size. All the *r*-values are statistically significant with $p \leq 10^{-5}$.

the other hand, we obtained a very high r-value (0.889) for archaeal Topo 1 which is nearly double that (0.464) of its eukaryal counterpart ($p<0.01$). Furthermore, the r-value of archaeal RFC3 was also significantly higher ($p<0.01$) than that of eukaryal RFC3. These suggest that archaeal Topo 1 and RFC3 evolved with PCNA in a more coordinated manner than their eukaryal counterparts. These results clearly indicate that the evolution of interacting proteins may be significantly different along different lineages. However, it should be mentioned that in some cases (for example, Ligase 1, MLH 1), the computed r-values are significantly high both in archaea and eukarya as is evident from Table 1. In these cases, the differences in the r-values are statistically insignificant and hence indicate negligible amount of difference in their coevolution in two different lineages. Furthermore, the r-value of XPG is nearly double in archaeal lineage than eukarya and the difference is statistically significant (Supplementary Table S5). The above results show that there exists a significant difference in the r-values of archaea and eukarya for some of the partners, while for the others, the differences are not significant. Thus, we can infer that there is a possibility of different order of structural and functional constrained working in different lineages to shape the correlated evolution of interacting partners.

When we took RFC3 and PCNA sequences from both archaeal and eukaryal species and calculated the r-value for this combined set; we obtained a very low r-value (0.202). As mentioned previously, in the present study we calculated the r-values for archaea and eukarya independently. Interestingly, we obtained a high r-value (0.745) for RFC3 in archaeal lineage whereas in the case of eukarya, we still obtained a negligible correlation coefficient value (0.235). The results indicate that the archaeal PCNA evolved in a coordinated way with its interacting partner RFC3. On the other hand, the eukaryal counterparts do not have a signature of correlated evolution. The above results again indicate that the coordinated evolution of the interacting proteins may be different for different lineages. We also observed that all the interacting partners do not always coevolve.

It would be interesting to construct the phylogenetic trees for PCNA and its ten interacting partners to get insight of their clustering feature. The bootstrapped phylogenetic trees are shown in Supplementary Figure 1 to 11. It is clear from Figure 1 that eukaryal PCNAs do form a single cluster. It has been already mentioned that seven interacting partners (DNA Ligase 1, DNA Polymerase delta, DNA Topoisomerase 1, DNA Topoisomerase 2, Flap endonuclease 1, MLH 1 and Uracil DNA glycosylase) among the ten showed positive high r-values ($r > 0.60$). We also observed similar trend of phylogenetic trees for the above-mentioned seven interacting partners as is evident from Supplementary Figure 2 to 8.

The two interacting partners RFC3 and XPG which showed low negative r-values (almost no correlation) indicate no evidence for coevolution of them with PCNA. On the other hand, WRN had comparatively lower r-value. The phylogenetic tree of WRN also clearly supports (Supplementary Figure 11) the low r-value. The striking difference of WRN with PCNA is that in the case of WRN, the two archaeal species *Methanococcus maripaludis* and *Methanosarcina acetivorans* fall within eukaryal lineage. On the other hand, the phylogenetic trees of RFC3 and XPG do not have any distinct difference in the clustering pattern of branches with PCNA phylogenetic tree. However, there are also differences in the arrangements within eukaryal kingdom. For example, *Oryza sativa* and *Arabidopsis thaliana* did not cluster together in the case of RFC3 and XPG.

The above study shows that in contrast to the expected coevolution of a protein with all of its interacting partners,

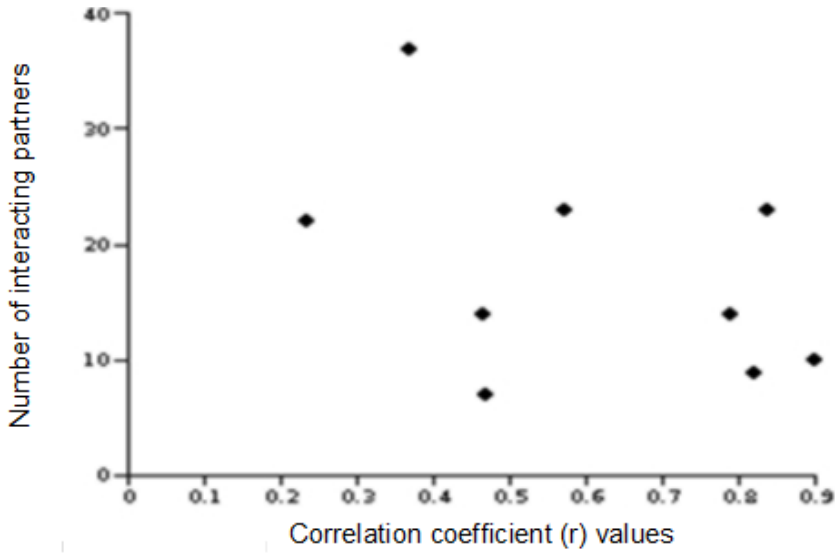


Figure 1. Relation between correlation coefficient values (*r*) and number of interacting partners.

PCNA interacting partners do not always coevolved with PCNA. It further indicates that the coordinated evolution of interacting partners is different for different lineages. Seven among the ten interacting partners having significantly high positive r-values, indicate the coevolution among PCNA and interacting partners, whereas the rest three do not have any signature of the coevolution. The 'entire sequence' approach used in our study deals with the pair wise distance matrix calculation of the alignment of the whole sequence. To understand the underlying reasons for wide variations in r-values, the important effect of cascading interactions and multiple interactions on the interacting protein partners is necessary to be addressed. A protein having multiple interacting partners (proteins) may exhibit different evolutionary pressures exerted by different interacting partners. Another probable reason for different order of evolutionary pressures may be following. Each of the interacting partners may also have interactions with other cellular proteins. For example the protein A may have interacting partners A1, A2 and A3. Again the protein A1 may have three partners A, A11 and A12. Each of the three proteins (A, A11 and A12) would provide structural and functional constraints to A1. Thus, when we consider the coevolution of A and A1 it is just not a pair of interactions (A and A1). Actually it is a cascading effect of coordinated pressures that ultimately develop the shape of so-called coevolution. For example, WRN, an interacting partner of PCNA has a large number of interacting partners, viz. P53, RAD52, RAD51, SUMO-1, Topo 2, RPA, etc. Furthermore, the nature and magnitude of pressure should depend on the functional importance of the complex and also on the number of interacting partners of each of the A1, A2 and A3. The pathway where the interacting proteins are involved may also be a determining factor. We have estimated the number of interacting partners of each of the ten interacting proteins of PCNA of eukarayal dataset using string database [http://string.embl.de], but failed to observe any direct correlation between r-values and number of protein partners (Figure 1). We took a stringent cut-off value of 0.9 score of STRING database to include number of interacting partners for each of the proteins. However, we did not observed any significant dependency of correlation coefficient values of PCNA interacting partners with their number of interacting partners. On the other hand, the significant point is that the interacting partners are widely involved in diverse kinds of biological pathways. The different pathways may impose different order of sequence-structure-functional constraint throughout the evolution.

Therefore, it would be very intriguing to understand whether there is any specific signature correlating the nature of coadaptation with the percentage of disorder region of the interacting partners. We identified the percentage of disorder regions of eukaryal and archaeal PCNA and the interacting proteins and the percentage of disorder regions are listed in Table 3.

We have found that all the PCNAs for the species mentioned in Table 3 had very lower percentage of disorder regions (data not shown here). On the other hand, the percentages of disorder regions of interacting proteins varied. We further classified the values (disorder region's percentages and r-values) into three groups-higher ($r \geq 0.6$ indicated as 1 (Table 4), lower ($r < 0.3$ indicated as -1), not determining ($r \geq 0.30$ and $r < 0.6$ indicated as 0). Based on this classification, using the data of Table 3 and the r-value listed in Table 2, we have

Table 3. Disorder percentage of ten PCNA interacting partners in Eukarya and Archaea lineage.

Interacting partner	Eukarya		Archaea		
	H.sapiens	***S. cerevisiae***	***A. pernix***	***M. jannaschii***	***P. abayssi***
Pold	9.382	11.704	21.018	13.636	17.932
Ligase 1	44.070	32.361	14.378	16.754	14.311
Fen1	13.684	15.707	13.217	12.883	13.994
Topo 1	34.771	40.702	12.630	21.813	16.350
Topo 2	32.819	26.102	11.429	2.989	7.068
RFC3	13.483	6.176	4.335	7.813	3.927
MLH 1	29.101	29.519	90.518	14.355	10.702
XPG	62.479	57.592	12.251	12.883	13.120
Uracil	23.642	27.019	6.977	9.742	1.531
WRN	34.008	39.529	13.390	7.729	10.197

Table 4. Relationship of disorder and *r* value . Higher disorder taken as 1 and lower disorder taken as -1.

Interacting partners	Eukarya			Archaea		
	PCNA disorder	**Disorder**	***r* value**	**PCNA disorder**	**Disorder**	***r* value**
Pold	-1	-1	1	-1	-1	1
Ligase 1	-1	1	1	-1	-1	1
Fen1	-1	-1	1	-1	-1	0
Topo 1	-1	1	0	-1	-1	1
Topo 2	-1	1	0	-1	-1	0
RFC3	-1	-1	-1	-1	-1	1
MLH 1	-1	1	1	-1	-1	1
XPG	-1	1	0	-1	-1	1
Uracil	-1	1	0	X	X	X
WRN	-1	1	0	X	X	X

Table 5. Representation of Disorder and *r*-value in Eukarya and Archaea.

Disorder of interacting partner	PCNA	*r*-value	Number of observation	
			Eukarya	**Archaea**
1	-1	0	5	0
1	-1	1	2	0
-1	-1	1	2	6
-1	-1	-1	1	0
-1	-1	0	0	2

further derived Table 4.

Table 4 shows some interesting observations which is again tabulated below in a derived Table 5. The predominant are - higher disorder (with higher percentage of disorder region) proteins which when interacted with PCNA (lower percentage of disorder region) give lower r- values (5 cases in eukarya). Lower disorder (with lower percentage of disorder region) proteins interact with PCNA (lower disorder) and give higher r-values (2 in eukarya and 6 in archaea). There are exceptions also indicating that the coevolution and coadaptation may have a relationship with percentage of disorder regions, however it alone cannot explain the wide range of r-values.

To understand the involvement of lineage specific selection pressures, we calculated the dn/ds ratio of both the eukarayal and archaeal dataset for PCNA as well as for each of its interacting partners. The basic idea behind

Table 6. Nonsynonymous(dn) to synonymous(ds) (dn/ds) ratio of Topo2 protein of archaeal organisms.

Organism	dn/ds	
	T. volcanium	***T. acidophilum***
Methanosarcina acetivorans	0.8431	0.8841
Methanosarcina mazei	1.2357	1.1750
Archaeoglobus fulgidus	0.7671	0.8871
Haloarculum marismortui	0.4643	0.4085
Methanothermobacter thermoautotrophicus	0.4783	0.8514
Methanocaladococcus jannaschi	1.0619	0.9334
Methanococcus maripaludis	0.9467	0.9784
Pyrococcus abyssi	1.2677	0.6475
Pyrococcus horikoshii	0.9783	0.9888
Pyrococcus furiosus	0.5977	0.7284

such study is that if the dn/ds ratio for any protein is >1, the protein is estimated to be under positive selection. If any PCNA interacting partners have dn/ds >1 they are not expected to coevolve. Similarly within the subset of interacting partner, if any organisms have dn/ds >1, then the organisms too are not expected to coevolve. In our study (Table 6 and for details see Supplementary Table S9) we found dn/ds >1 in the case of Topo2 in archaeal set for few organisms. Archaeal Topo2 r-value is comparatively lower than other interacting partners. Probably, this may be one of the reasons of its lower r-value. In the case of eukaryal Topo2, although the r-value is quite low (0.405) but we did not get any positive selection pressure in that protein set. Topo1 in eukaryal dataset shows low r-value but the dn/ds in this case was less than 1. RFC3 also shows a very low r-value in eukaryal dataset while dn/ds ratio did not give us any indication of positive selection pressure. Archaeal Fen1 dataset also showed low correlation coefficient value but only negative selection pressure existed. So, by estimating dn/ds ratio alone, enough clue of correlated evolution of PCNA interacting protein set was not gotten.

Moreover, existing literature suggests that the interacting partners of PCNA are involved in various functional pathways, viz, DNA Polymerase delta, Replication factor C3, DNA Ligase 1, DNA Topoisomerase 1, DNA Topoisomerase 2; are involved in DNA replication and repair, MLH1 in mismatch DNA repair, XPG endonuclease in nucleotide excision repair, WRN helicase in double strand breaks DNA repair and Uracil DNA glycosylase in base excision repair (Giovani and Ulrich, 2003). Thus, the interacting partners having involvement in a number of different functional pathways exhibit different orders of pressures to maintain their structural and functional integrity. Finally, we can say that evolutionary relationships of a protein with its multiple interacting partners (proteins) depend on several factors that need a future study.

In summary, the evolution of a protein having multiple interacting partners is governed by the structural and functional constraints imposed by its partners. The interacting partner proteins may have different order of controls on the protein which result in differences in their coevolutionary pattern. In addition, the present work shows that the natures of coevolution of the interacting proteins are different in case of the eukaryal and archaeal lineages. The possible structural and functional constraints and their possible influences have also been discussed. It has been observed that the percentage of order and percentage of disorder region of the interacting proteins appear as the most significant determinant of their coevolutionary pattern. However, we should mention that not any single constraint (percentage of order and disorder region of proteins) but a set of constraints like cascading effects of interaction of interacting partners, their functional constraints, etc. should also play important roles in shaping the coevolutionary nature of multiple-interacting proteins.

ACKNOWLEDGEMENTS

The authors sincerely thank Dr. U. Chaudhuri of the University of Calcutta for his suggestions. We also want to thank Mr. Ananyo Choudhuri for reading the Manuscript. We acknowledge the DIC, University of Calcutta for using the computational facility. S. Biswas thanks the University Grant Commission, India for providing fellowship.

REFERENCES

Altschuh D, Lesk AM, Bloomer AC, Klug A (1987). Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. J. Mol. Biol. 193:693-707.

Altschul SF, Gish W, Miller W, Mayers EW, Lipman DJ (1990) Basic local alignment search tool. J. Mol. Biol. 215:403-410.

Argos P (1988). An investigation of protein subunit and domain interfaces. Protein Eng. 2:101-113.

Atwell S, Ultsch M, Vos DMA, Wells AJ (1997). Structural plasticity in a remodeled protein-protein interface. Science 278:1125-1128.

Bork P, Jensen LJ, Mering CV, Ramani AK, Lee I, Marcotte EM (2004). Exploiting the co-evolution of interacting proteins to discover interaction specificity. Curr. Opin. Struct. Biol. 14:292-299.
Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradović Z(2002).Intrinsic disorder and protein function. Biochemistry 41:6573-6582.
Felsenstein J (2002). PHYLIP: Phylogeny inference package, version 3.6, Department of Genome Sciences, University of Washington, Seattle.
Giovanni M, Ulrich H (2003). Proliferating cell nuclear antigen (PCNA): a dancer with many partners. Cell Sci. 116:3051-3060.
Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE (2000). Co-evolution of proteins with their interaction partners. J. Mol. Biol. 299:283-293.
Goh CS, Cohen FE (2002). Co-evolutionary analysis reveals insights into protein-protein interactions. J. Mol. Biol. 324:177-192.
Grigoriev A (2003). On the number of protein-protein interactions in the yeast proteome. Nucleic Acids Res. 31:4157-4161.
Higgins D, Thompson J, Gibson T, Thompson JD, Higgins DG, Gibson TJ (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673-4680.
Hoskins J, Lovell S, Blundell TL (2006) An algorithm for predicting protein-protein interaction sites: Abnormally exposed amino acid residues and secondary structure elements. Protein Sci. 15:1017-1029.
Janin J, Miller S, Chothia C (1988) Surface, subunit interfaces and interior of oligomeric proteins. J. Mol. Biol. 204:155-164.
Jespers L, Lijnen HR, Vanwetswinkel S, Hoef VB, Brepoels K, Cohen D, Maeyer DM (1999) Guiding a docking mode by phage display: selection of correlated mutations at the staphylokinase-plasmin interface. J. Mol. Biol. 290:471-479.
Jones S, Thornton JM (1996) Principles of protein-protein interactions. Proc. Natl. Acad. Sci. USA 93:13-20.
Jonsson ZO, Hubscher U (1997). Proliferating cell nuclear antigen:more than a clamp for DNA polymerases. Bioessays 19:967-975.
Kim WK, Bolser DM, Park JH (2004). Large-scale co-evolution analysis of protein structural interlogues using the global protein structural interactome map (PSIMAP). Bioinformatics 20:1138-1150.
Mintseris J, Weng Z (2005) Principles of protein-protein interactions. Proc. Natl. Acad. Sci. USA 102:10930-10935.
Moyle WR, Campbell RK, Myers RV, Bernard MP, Han Y, Wang X (1994). Co-evolution of ligand-receptor pairs. Nature 368:251-255.
Ohta T (1995). Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. J. Mol. Evol. 40:56-63.
Pal C, Papp B, Lercher MJ (2006). An integrated view of protein evolution. Nat. Rev. Genet. 7:337-348.
Pazos F, Helmer-Citterich M, Ausiello G, Valencia A (1997) Correlated mutations contain information about protein-protein interaction. J. Mol. Biol. 271:511-523.
Pazos F, Juan D, Izarzugaza JM, Leon E, Valencia A (2008) Prediction of protein interaction based on similarity of phylogenetic trees. Methods Mol. Biol. 484:523-35.
Pazos F, Ranea JA, Juan D, Sternberg MJ (2005) Assessing Protein Co-evolution in the Context of the tree of life assists in the prediction of the interactome. J. Mol. Biol. 30; 1002-1015.
Pazos F, Valencia A (2001) Similarity of Phylogenetic trees as indicator of protein-protein interaction. Protein Eng. 14:609-614.
Pazos F, Valencia A (2008). Protein co-evolution, co-adaptation and interactions. EMBO J. 27:2648-2655.
Pereira-Leal JB, Levy ED, Teichmann SA (2006) The origins and evolution of functional modules: lessons from protein complexes. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 361:507-517.
Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992). Numerical Recipes in C: the Art of Scientific Computing. 2nd edition. Cambridge University Press, Cambridge.
Ramani AK, Marcotte EM (2003). Exploiting the co-evolution of interacting proteins to discover interaction specificity. J. Mol. Biol. 327:273-284.
Sato T, Yamanishi Y, Kanehisa M, Toh H (2005). The KEGG resource for deciphering the genome. Bioinformatics 21:3482-3489.
Spiegel RM (1972) Theory and Problems of Statistics in Si units, 1st edition, McGraw-Hill International Book Company, New York. pp. 264.
Tan SH, Zhang Z, Ng SK (2004) ADVICE: Automated Detection and Validation of Interaction by Co-Evolution. Nucleic Acids Res. 32:W69-W72.
Warbrick E (2000). The puzzle of PCNA's many partners. Bioessays 22:997-1006.

Supplementary Table S1A

Organism	PCNA	Interacting Protein Partner									
		RFC3	POLd	DNA Ligase I	TOPO1	TOPO2	FEN 1	XPG endonuclease	WRN helicase	MLH1	URACIL
Eukarya											
Homo sapiens	CAG46598	P40938	P49005	P18858	P11387	CAA48197	P39748	P28715	AAC41981	AAC50285	P13051
Mus musculus	P17918	Q8R323	CAA96567	AAH28287	BAA00950	BAA02076	A53730	P35689	AAC72359	NP_081086	NP_035807
Rattus norvegicus	AAH60570	AAH88281*	AAH79267*	NP_110482	AAD30137	CAA86496	AAH83630	XP_217387	XP_232510	NP_112315	XP_222272
Drosophila melanogaster	A34752	AAF63387	Q9W088	AAF47090	P30189	P15348	NP_523765	AAD50779	Q9VGI8	NP_477022	NP_573064
Caenorhabditis elegans	O02115	NP_502517	Q19366#	NP_741625	CAA65537	NP_496536	NP_491168	AAB96723@	AAM26298	Q9XU10@	NP_499560
Saccharomyces cerevisiae	AAB31034	NP_014109	P46957	CAA91582	P04786	AAB36610	NP_012809	P07276	NP_013915	P38920	A31425
Schizosaccharomyces pombe	CAA38636	O14003	P87324	CAA28754	P07799	NP_595805	NP_594972@	P28706	CAA70577^	Q9P7W6	AAD51974
Arabidopsis thaliana	AAM63900	BAB67768$	O48520	CAA66599	BAB08548	P30182	AAC13596	Q9ATY5	AAG50580^	AAK25988#	NP_188493
Oryza sativa	P17070	Q9FXT5	Q9LRE5	NP_922089^	Q84ZL5^	XP_467311	BAA36171	BAB72003	XP_479556^	BAB89000#	XP_474316

*, Predicted: #, probable; ^, putative; $, like; @, hypothetical; &, similar; !@,conserve hypothetical; =, related sequence type.

Supplementary Table S1B

Organism	PCNA	Interacting Protein Partner									
		RFC3	POLd	DNA Ligase I	TOPO1	TOPO2	FEN 1	XPG endonuclease	WRN helicase	MLH1	URACIL
Archaea											
Aeropyrum pernix	NP_147232	Q9YBS7	BAA80016@	NP_147713	Q9YB01	Q9YE67	NP_146975	Q9YFY5	Q9YFQ8#	BAA80709@	NP_147220@
Archaeoglobus fulgidus	NP_069171	O28219	E69473#	NP_069457	O28469	O29322	NP_069102	O29975	AAB90094=	NP_069865	NP_071102@
Methanocaldococcus jannaschii	NP_247218	Q58817	NP_247686	NP_247139	Q59046	Q57815	NP_248448	Q58839	AAB98279=	NP_247618@	NP_248433@
Methanosarcina acetovorans	NP_615084	NP_615630	NP_615011	NP_615688	Q8TMY4	Q8TQF8	NP_618874	Q8TIY5	AAM07847	NP_615486	NP_618469
Methanosarcina mazei	NP_633421	Q8PVY4	NP_633369	NP_633919	Q8PSK3	Q8PUB7	NP_632930	Q8PYF6	AAM30913	NP_633706	NP_632510@
Pyrococcus abyssi	Q9UYX8	Q9V2G4	Q9V2F3	CAC20743	Q9UYS8	Q9V134	NP_126423	Q9V0P9	CAB49731=	NP_127265	NP_126375@
Pyrococcus furiosus	AAL81107	NP_577822	P81412	NP_579364	O73954	Q8U0K9	AAD01514	O93634	NP_577782#	NP_578203	NP_579114@
Sulfolobus solfataricus	NP_341944	AAK41065	AAK42021!@	NP_341745	NP_342400	Q97ZE9	NP_341735	Q980U8	AAK41239	AAK42909!@	NP_343647@
Thermoplasma volcanium	BAB60230	Q977Z9	NP_111891#	NP_111756	NP_110538	BAB59299@	BAB59701	Q97B98	NP_111333*	BAB59815	NP_111346

*, Predicted; #, probable; ^, putative; $, like; @, hypothetical; &, similar; !@, conserve hypothetical; =, related sequence type.

Supplementary Table S2A

Organism	PCNA	Interacting Protein Partner									
		RF-C	POLdelta	DNA Ligase I	TOPO1	TOPO2	FEN 1	XPG endonuclease	WRN helicase	MLH	URACIL
Eukarya											
Homo sapiens	CAG46598	P40938	P49005	P18858	P11387	CAA48197	P39748	P28715	AAC41981	AAC50285	P13051
Mus musculus	P17918	Q8R323	CAA96567	AAH28287	BAA00950	BAA02076	A53730	P35689	AAC72359	NP_081086	NP_035807
Rattus norvegicus	AAH60570	-	-	NP_110482	AAD30137	CAA86496	AAH83630	XP_217387	XP_232510	NP_112315	XP_222272
Drosophila melanogaster	A34752	AAF63387	Q9W088	AAF47090	P30189	P15348	NP_523765	AAD50779	Q9VGI8	NP_477022	NP_573064
Caenorhabditis elegans	O02115	NP_502517	-	NP_741625	CAA65537	NP_496536	NP_491168	-	AAM26298	-	NP_499560
Saccharomyces cerevisiae	AAB31034	NP_014109	P46957	CAA91582	P04786	AAB36610	NP_012809	P07276	NP_013915	P38920	A31425
Schizosaccharomyces pombe	CAA38636	O14003	P87324	CAA28754	P07799	NP_595805	-	P28706	-	Q9P7W6	AAD51974
Arabidopsis thaliana	AAM63900	-	O48520	CAA66599	BAB08548	P30182	AAC13596	Q9ATY5	-	-	NP_188493
Oryza sativa	P17070	Q9FXT5	Q9LRE5	-	-	XP_467311	BAA36171	BAB72003	-	-	XP_474316
Xenopus laevis	P18248	-	O93610	-	P41512	-	O57351	-	-	-	-
Cryptosporidium hominis	Q5CJE0	EAL36856	-	-	EAL38033	EAL35251	EAL36849	-	-	-	-
Gallus gallus	Q9DEA3	NP_001006276	-	-	BAA19101	O42130	Q90YB0	-	-	-	-
Plasmodium falciparum	P61074	AAG37985	-	CAD52175	CAA58716	P41001	-	-	-	-	-
Daucus carota	Q00268	-		-	Q9XGL1	-	-	-	-	-	-
Danio rerio	Q9PTP1	-	AAH66617	-	-	AAH86970	-	-	-	-	-
Encephalitozoon cuniculi	Q8SRV9	-	Q8SQN5	-	-	NP_584718	-	-	-	-	-
Cryptosporidium parvum	Q5CUB6	-	-	EAK88972	EAK90633	EAK87659	-	-	-	-	-
Pisum sativum	CAA76392	-	-	-	CAA74890	CAA74891	-	-	-	-	-
Nicotiana tabacum	CAA77062	-	-	-	AAK69776	AAN85208	-	-	-	-	-

Supplementary Table S2B

#Organism	PCNA	Interacting Protein Partner					
		RF-C	POLdelta	DNA Ligase I	TOPO1	TOPO2	FEN 1
Aeropyrum pernix	NP_147232	Q9YBS7	-	NP_147713	Q9YB01	Q9YE67	NP_146975
Archaeoglobus fulgidus	NP_069171	O28219	E69473	NP_069457	O28469	O29322	NP_069102
Methanocaldococcus jannaschii	NP_247218	Q58817	NP_247686	NP_247139	Q59046	Q57815	NP_248448
Methanosarcina acetovorans	NP_615084	NP_615630	NP_615011	NP_615688	Q8TMY4	Q8TQF8	NP_618874
Methanosarcina mazei	NP_633421	Q8PVY4	NP_633369	NP_633919	Q8PSK3	Q8PUB7	NP_632930
Pyrococcus abyssi	Q9UYX8	Q9V2G4	Q9V2F3	CAC20743	Q9UYS8	Q9V134	NP_126423
Pyrococcus furiosus	AAL81107	NP_577822	P81412	NP_579364	O73954	Q8U0K9	AAD01514

#This set of organisms sequence taken from NCBI database.

Supplementary Table S2B. Contd

Sulfolobus solfataricus	NP_341944	AAK41065	-	NP_341745	NP_342400	Q97ZE9	NP_341735
Thermoplasma volcanium	BAB60230	Q977Z9	NP_111891	NP_111756	NP_110538	BAB59299	BAB59701
Methanococcus maripaludis	CAF31267	Q6M044	CAF29564	Q6LYM1	Q6LYN4	NP_988557	CAF30869
Methanothermobacter thermoautotrophicus	O27367	O26343	O27456	Q50566	O27661	NP_276143	O27670
Pyrococcus horikoshii	O58398	O57852	O57863	NP_143476	O58356	BAA30675	O50123
Sulfolobus tokodaii	Q975N2	Q975D3	-	NP_376074	NP_377148	BAB66339	Q976H6
Thermoplasma acidophilum	CAC12046	Q9HI47	Q9HLK5	Q9HJ26	Q9HM08	CAC11245	Q9HJD4
Haloarculum marismortui	CAB93143	AAV47358	AAV47482	-	AAV46558	AAV45488	AAV45115
Themococcus fumicolans	CAB59006	-	-	CAC21199	-	-	-

Supplementary Table S2B. Contd.

*Organism	PCNA	MLH 1	Uracil	WRN	XPG
Anopheles gambiae	Q7Q0Q0	Q7QIY1	-	-	-
Bombyx mori	BGIBMGA010906-PA	BGIBMGA012027-PA	-	-	Bmb016974
Bos taurus	ENSBTAP00000007967	ENSBTAP00000022288	ENSBTAP00000026445	ENSBTAP00000045836	ENSBTAP00000000071
Canis familiaris	ENSCAFP00000009045	ENSCAFP00000007136	ENSCAFP00000016407	ENSCAFP00000019003	
Danio rerio	ENSDARP00000070780	ENSDARP00000034180	ENSDARP00000062358	ENSDARP00000072210	ENSDARP00000004016
Drosophila melanogaster	P17917	Q9V380	Fbgn0038490	Fbgn0011802	Fbgn0004584
Encephalitozoon cuniculi	Q8SRV9	Q8SS00	Q8SR60	Q8SQJ7	
Eremothicium gossypii	Q75B81	Q755L3	Q756E0	Q759G7	Q74ZJ5
Gallus gallus	ENSGALP00000040305	ENSGALP00000019676	ENSGALP00000033675	ENSGALP00000027374	ENSGALP00000005724
Homo sapiens	ENSP00000368458	ENSP00000231790	ENSP00000242576	ENSP00000351886	ENSP00000305480
Macaca mulatta	ENSMMUP00000017436	ENSMMUP00000029671	ENSMMUP00000009131	ENSMMUP00000016596	ENSMMUP00000008754
Mus musculus	ENSMUSP00000028817	ENSMUSP00000035079	ENSMUSP00000031587	ENSMUSP00000086312	ENSMUSP00000025651
Neurospora crassa	Q7SF71	Q7SA79	Q7SG58	Q872I5	Q7SC91
Pan troglodytes	-	-	ENSPTRP00000009190	ENSPTRP00000042003	ENSPTRP00000006454
Rattus norvegicus	ENSRNOP00000028887	ENSRNOP00000043097	ENSRNOP00000000872	ENSRNOP00000044668	ENSRNOP00000027842
Saccharomyces cerevisiae	P15873	P38920	Q06244	P53115	Q02825
Schizosaccharomyces pombe	Q03392	Q9P7W6	O74834	O13682	O74908
Xenopus tropicalis	ENSXETP00000017963	ENSXETP00000010905	ENSXETP00000048248	ENSXETP00000032370	ENSXETP00000014663

*This set of Organisms sequence taken from Orthodb database.

Supplementary Table S3. Z-values obtained from two-tailed test to predict whether any two calculated r-values are satistically significant or not.

Protein name	Combined all with hypotheticals								
	POLd	**TOPO1**	**TOPO2**	**FEN1**	**RFC3**	**WRN**	**XPG**	**URACIL**	**MLH 1**
LIGASE 1	0.9	-1.06	-0.73	-1.87	9.09	5	9.39	2.19	2.08
POLd		-1.97	-1.63	-2.78	8.19	4.09	8.48	1.28	1.18
TOPO1			0.34	-0.8	10.16	6.07	10.46	3.26	3.15
TOPO2				-1.15	9.82	5.72	10.11	2.91	2.81
FEN1					10.96	6.87	11.26	4.06	3.95
RFC3						-4.09	0.3	-6.9	-7
WRN							4.39	-2.81	-2.91
XPG								-7.2	-7.3
URACIL									-0.1

The r-values are statistically significantly different at the 0.10%, 0.05% and 0.01% level if the |Z| values are greater than 1.65, 1.96 and 2.58, respectively. The results are given for the r-values obtained using the combined sequences (dataset 1).

Supplementary Table S4. Z-values obtained from two-tailed test to predict whether any two calculated r-values are satistically significant or not.

Protein name	Combined all without hypotheticals								
	POLd	**TOPO1**	**TOPO2**	**FEN1**	**RFC3**	**WRN**	**XPG**	**URACIL**	**MLH 1**
LIGASE 1	-3.16	0.05	-2.21	-2.03	6.11	2.07	8.27	1.34	0.84
POLd		3.2	1.39	1.53	8.04	4.12	9.79	3.75	3.43
TOPO1			-2.26	-2.09	6.06	2.03	8.22	1.3	0.8
TOPO2				0.17	8.24	3.46	10.48	3	2.61
FEN1					8.08	3.35	10.26	2.87	2.47
RFC3						-1.89	1.85	-3.33	-4.13
WRN							3.12	-0.82	-1.27
XPG								-4.83	-5.75
URACIL									-0.48

The r-values are statistically significantly different at the 0.10, 0.05 and 0.01% level if the |Z| values are greater than 1.65, 1.96 and 2.58, respectively. The results are given for the r-values obtained using the combined sequences (dataset 2).

Supplementary Table S5. Z-values obtained from two-tailed test to predict whether any two calculated r-values are statistically significant or not.

Protein name	Z value
Without hypothetical Eukarya and Archaea	
LIGASE 1	0.28
POLd	3.03
TOPO1	-5.82
TOPO2	-1.45
FEN1	2.37
RFC3	-4.04
MLH 1	0.29
URACIL	N/A
XPG	-2
WRN	N/A

Supplementary Table S5. Contd.

Without hypothetical Eukarya and Archaea (0.1%>1.65)	
LIGASE 1	
POLd	3.03
TOPO1	-5.82
TOPO2	
FEN1	2.38
RFC3	-4.04
MLH 1	
URACIL	N/A
XPG	-2
WRN	N/A

The r-values are statistically significantly different at the 0.10, 0.05 and 0.01% level if the |Z| values are greater than 1.65, 1.96 and 2.58, respectively. The results are given for the r-values obtained using the without hypotheticals Eukarya and Archaea (dataset 3).

Supplementary Table S6. Z-values obtained from two-tailed test to predict whether any two calculated r-values are satistically significant or not.

Protein name	Eukarya without hypothetical						
	POLd	**TOPO1**	**TOPO2**	**FEN1**	**RFC3**	**XPG**	**MLH 1**
LIGASE 1	-1.43	3.58	4.04	0.43	4.15	2.91	-0.78
POLd		5.32	5.81	2	5.59	4.14	0.17
TOPO1			0.57	-3.57	1.46	0.41	-2.97
TOPO2				-4.13	1.07	0.08	-3.23
FEN1					4.12	2.75	-1.08
RFC3						-0.68	-3.55
WRN						-0.26	-2.65
XPG							-2.82
URACIL							-3.77

The r-values are statistically significantly different at the 0.10, 0.05 and 0.01% level if the |Z| values are greater than 1.65, 1.96 and 2.58, respectively. The results are given for the r-values obtained using the without hypotheticals eukaryal sequences (dataset 3).

Supplementary Table S7. Z-values obtained from two-tailed test to predict whether any two calculated r-values are satistically significant or not.

Protein name	Archaea without hypothetical						
	POLd	**TOPO1**	**TOPO2**	**FEN1**	**RFC3**	**XPG**	**MLH 1**
LIGASE 1	0.39	-3.22	2.68	2.23	-0.33	0.54	-0.66
POLd		3.2	1.95	1.56	0.67	0.22	0.83
TOPO1			5.9	5.45	2.89	2.79	0.82
TOPO2				-0.45	-3	-1.3	-1.88
FEN1					-2.55	-1.02	-1.68
RFC3						0.76	-0.5
XPG							-0.91

The r-values are statistically significantly different at the 0.10, 0.05 and 0.01% level if the |Z| values are greater than 1.65, 1.96 and 2.58, respectively. The results are given for the r-values obtained using the without hypotheticals archaeal sequences (dataset 3).

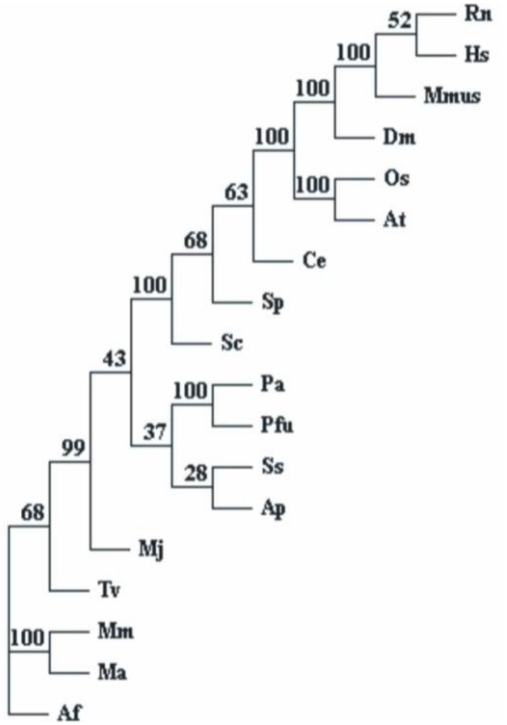
Supplementary Table S8. Z-values obtained from two-tailed test to predict whether any two calculated r-values are satistically significant or not.

Protein name	Combined with and without hypothetical
	Z value
LIGASE 1	-0.13
POLd	3.81
TOPO1	-1.22
TOPO2	1.44
FEN1	0.15
RFC3	2.11
MLH 1	0.64
URACIL	0.11
XPG	0.5
WRN	0.83

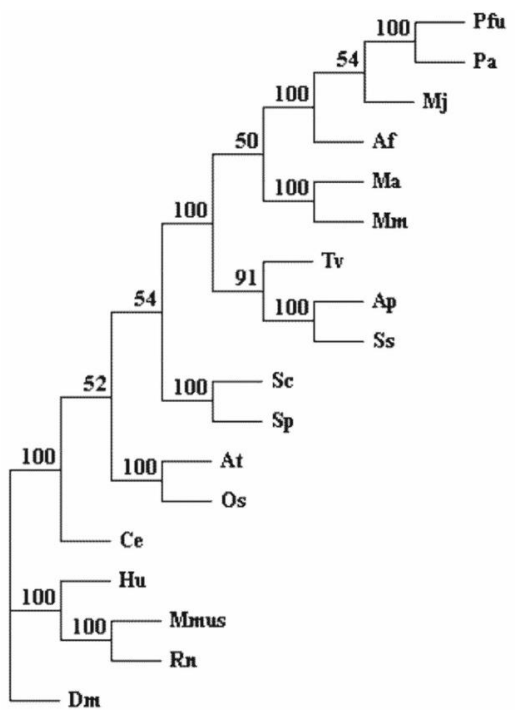
The r-values are statistically significantly different at the 0.10%, 0.05% and 0.01% level if the |Z| values are greater than 1.65, 1.96 and 2.58, respectively.The results are given for the r-values obtained using the combined with hypotheticals and comined without hypotheticals sequences (dataset 1 and dataset 2).

Supplementary Table S9.

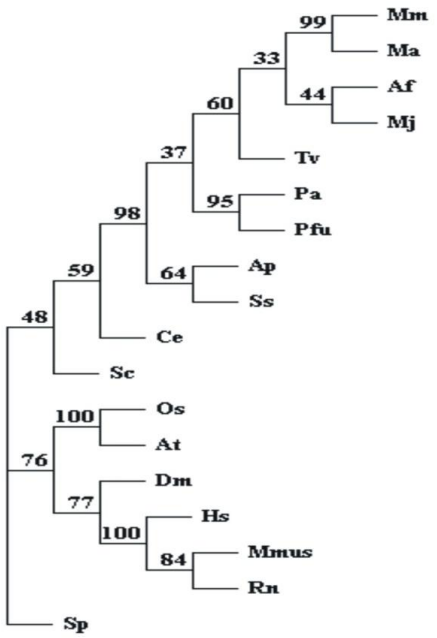
	Ma_Q8TMY4	Mm_Q8PSK3	hmAAV46558	Af_O28469	Tv_110538	taQ9HM08	Mj_Q59046	mpQ6LYN4	Pf_O73954	phO58356	Pa_Q9UYS8	mtO27661
Ma_Q8TMY4												
Mm_Q8PSK3	0.0592											
hmAAV46558	0.0995	0.0994										
Af_O28469	0.1038	0.1281	0.114									
Tv_110538	0.2806	0.3151	0.1436	0.1315								
taQ9HM08	0.1447	0.2264	0.1459	0.2235	0.0537							
Mj_Q59046	0.175	0.1844	0.1851	0.1689	0.2182	0.1889						
mpQ6LYN4	0.1646	0.1616	0.1778	0.1648	0.1901	0.1852	0.0582					
Pf_O73954	0.3232	0.2744	0.1824	0.1742	0.3781	0.3571	0.0948	0.1077				
phO58356	0.1773	0.2728	0.1845	0.3275	0.3082	0.4206	0.0997	0.1073	0.0164			
Pa_Q9UYS8	0.3973	0.3681	0.175	0.1771	0.4271	0.2697	0.1028	0.1138	0.0253	0.0304		
mtO27661	0.1757	0.1766	0.1743	0.2899	0.1987	0.2159	0.1309	0.1328	0.1298	0.1276	0.1254	



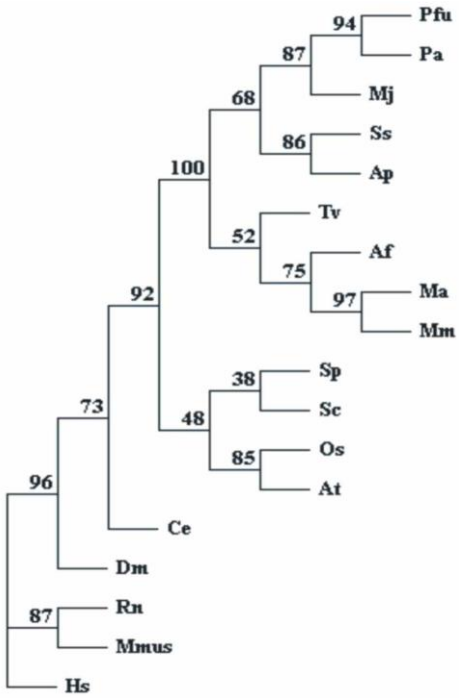
Supplementary Figure 1. Bootstrapped phylogenetic tree of PCNA. Ap = *Aeropyrum pernix*; Af = *Archaeoglobus fulgidus*; At = *Arabidopsis thaliana* ;Ce = *Caenorhabditis elegans*; Dm = *Drosophila melanogaster*; Hu = *Homo sapiens*; Ma = *Methanosarcina acetovorans*; Mj = *Methanocaldococcus jannaschii*; Mm = *Methanosarcina mazei*; Mmus = *Mus musculus*; Os = *Oryza sativa*; Pa = *Pyrococcus abyssi*; Pfu = *Pyrococcus furiosus*; Rn = *Rattus norvegicus*; Sc = *Saccharomyces cerevisiae*; Sp = *Schizosaccharomyces pombe*; Ss = *Sulfolobus solfataricus*; Tv = *Thermoplasma volcanium.* The number in each node indicates the confidence value of that branch after bootstrapping the phylogenetic tree.



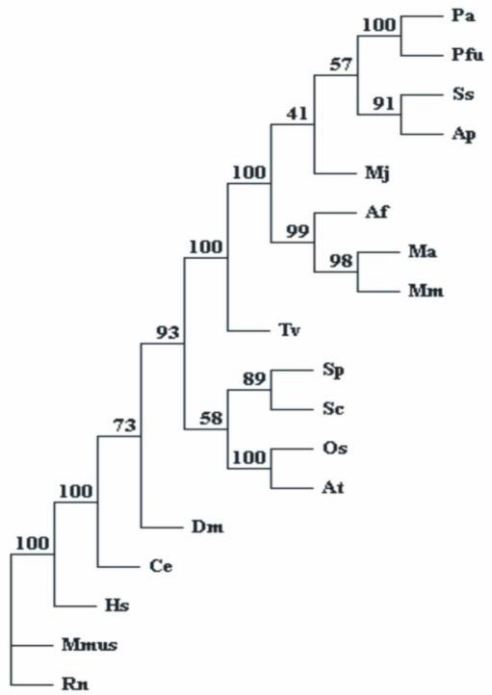
Supplementary Figure 2. Bootstrapped phylogenetic tree of Ligase 1. Ap = *Aeropyrum pernix*; Af = *Archaeoglobus fulgidus*; At = *Arabidopsis thaliana*; Ce = *Caenorhabditis elegans*; Dm = *Drosophila melanogaster*; Hu = *Homo sapiens*; Ma = *Methanosarcina acetovorans*; Mj = *Methanocaldococcus jannaschii*; Mm = *Methanosarcina mazei*; Mmus = *Mus musculus*; Os = *Oryza sativa*; Pa = *Pyrococcus abyssi*; Pfu = *Pyrococcus furiosus*; Rn = *Rattus norvegicus*; Sc = *Saccharomyces cerevisiae*; Sp = *Schizosaccharomyces pombe*; Ss = *Sulfolobus solfataricus*; Tv = *Thermoplasma volcanium.* The number in each node indicates the confidence value of that branch after bootstrapping the phylogenetic tree.



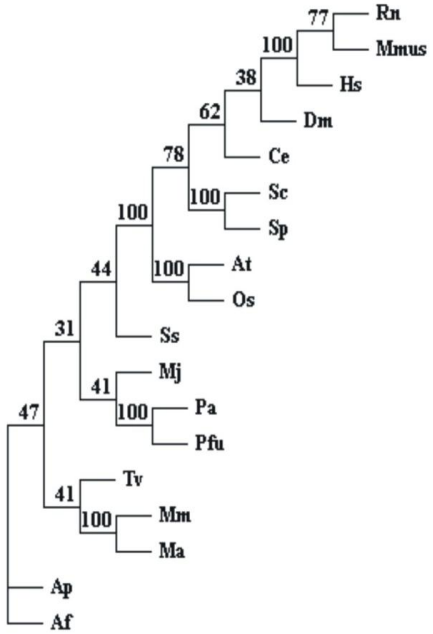
Supplementary Figure 3. Bootstrapped Phylogenetic tree of Polymerase delta. Ap = *Aeropyrum pernix*;Af = *Archaeoglobus fulgidus*;At = *Arabidopsis thaliana* ;Ce = *Caenorhabditis elegans* ;Dm = *Drosophila melanogaster*;Hu = *Homo sapiens*;Ma = *Methanosarcina acetovorans*;Mj = *Methanocaldococcus jannaschii*;Mm = *Methanosarcina mazei*;Mmus = *Mus musculus*;Os = *Oryza sativa*;Pa = *Pyrococcus abyssi*;Pfu = *Pyrococcus furiosus*;Rn = *Rattus norvegicus*;Sc = *Saccharomyces cerevisiae*;Sp = *Schizosaccharomyces pombe*;Ss = *Sulfolobus solfataricus*;Tv = *Thermoplasma volcanium.* The number in each node indicatates the confidence value of that branch after bootstrapping the phylogenetic tree.



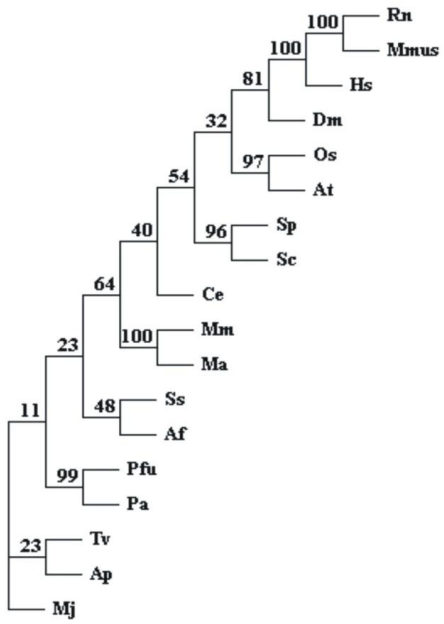
Supplementary Figure 4. Bootstrapped Phylogenetic tree of Topoisomerase 1. Ap = *Aeropyrum pernix*;Af = *Archaeoglobus fulgidus*;At = *Arabidopsis thaliana* ;Ce = *Caenorhabditis elegans* ;Dm = *Drosophila melanogaster*;Hu = *Homo sapiens*;Ma = *Methanosarcina acetovorans*;Mj = *Methanocaldococcus jannaschii*;Mm = *Methanosarcina mazei*;Mmus = *Mus musculus*;Os = *Oryza sativa*;Pa = *Pyrococcus abyssi*;Pfu = *Pyrococcus furiosus*;Rn = *Rattus norvegicus*;Sc = *Saccharomyces cerevisiae*;Sp = *Schizosaccharomyces pombe*;Ss = *Sulfolobus solfataricus*;Tv = *Thermoplasma volcanium.* The number in each node indicatates the confidence value of that branch after bootstrapping the phylogenetic tree.



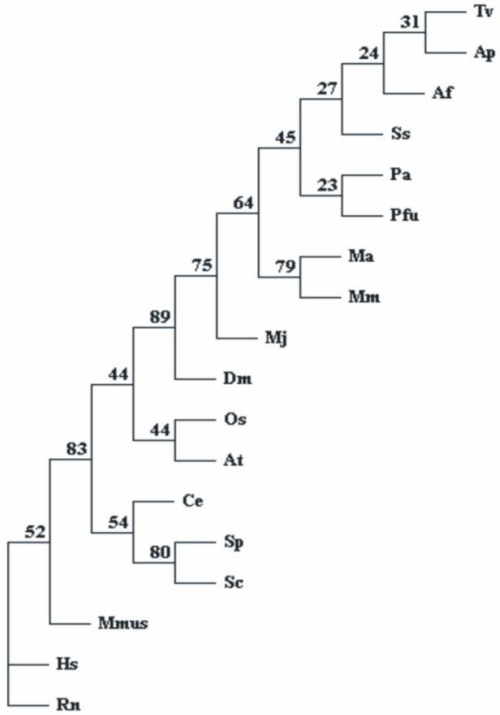
Supplementary Figure 5. Bootstrapped Phylogenetic tree of Topoisomerase 2. Ap = *Aeropyrum pernix*;Af = *Archaeoglobus fulgidus*;At = *Arabidopsis thaliana* ;Ce = *Caenorhabditis elegans* ;Dm = *Drosophila melanogaster*;Hu = *Homo sapiens*;Ma = *Methanosarcina acetovorans*;Mj = *Methanocaldococcus jannaschii*;Mm = *Methanosarcina mazei*;Mmus = *Mus musculus*;Os = *Oryza sativa*;Pa = *Pyrococcus abyssi*;Pfu = *Pyrococcus furiosus*;Rn = *Rattus norvegicus*;Sc = *Saccharomyces cerevisiae*;Sp = *Schizosaccharomyces pombe*;Ss = *Sulfolobus solfataricus*;Tv = *Thermoplasma volcanium.* The number in each node indicatates the confidence value of that branch after bootstrapping the phylogenetic tree.



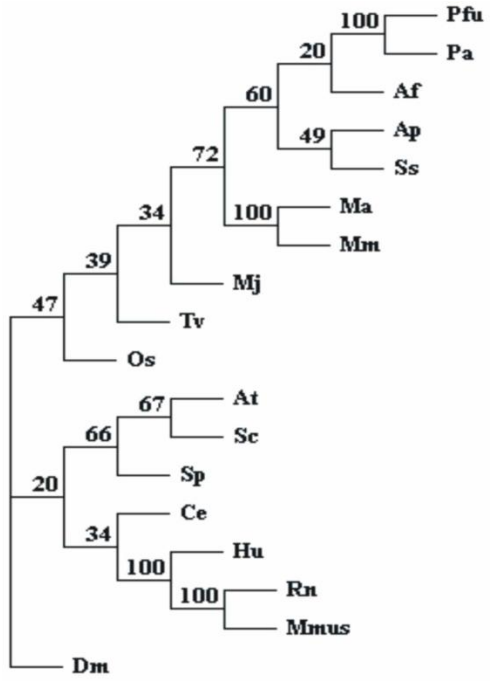
Supplementary Figure 6. Bootstrapped Phylogenetic tree of Fen 1. Ap = *Aeropyrum pernix*;Af = *Archaeoglobus fulgidus*;At = *Arabidopsis thaliana* ;Ce = *Caenorhabditis elegans* ;Dm = *Drosophila melanogaster*;Hu = *Homo sapiens*;Ma = *Methanosarcina acetovorans*;Mj = *Methanocaldococcus jannaschii*;Mm = *Methanosarcina mazei*;Mmus = *Mus musculus*;Os = *Oryza sativa*;Pa = *Pyrococcus abyssi*;Pfu = *Pyrococcus furiosus*;Rn = *Rattus norvegicus*;Sc = *Saccharomyces cerevisiae*;Sp = *Schizosaccharomyces pombe*;Ss = *Sulfolobus solfataricus*;Tv = *Thermoplasma volcanium.* The number in each node indicatates the confidence value of that branch after bootstrapping the phylogenetic tree.



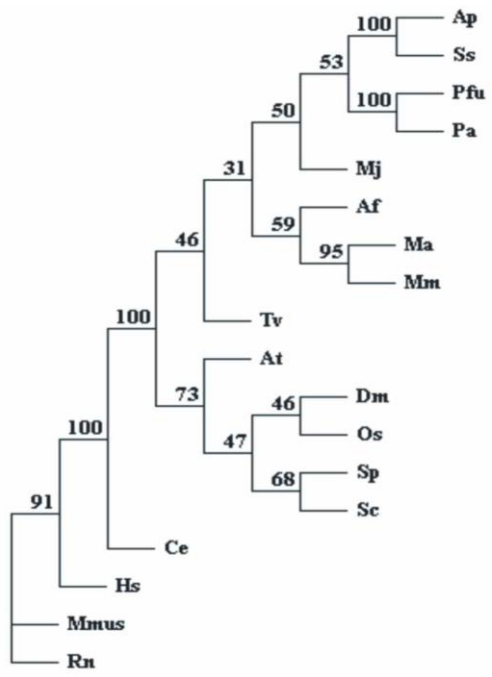
Supplementary Figure 7. Bootstrapped Phylogenetic tree of MLH1. Ap = *Aeropyrum pernix*;Af = *Archaeoglobus fulgidus*;At = *Arabidopsis thaliana* ;Ce = *Caenorhabditis elegans* ;Dm = *Drosophila melanogaster*;Hu = *Homo sapiens*;Ma = *Methanosarcina acetovorans*;Mj = *Methanocaldococcus jannaschii*;Mm = *Methanosarcina mazei*;Mmus = *Mus musculus*;Os = *Oryza sativa*;Pa = *Pyrococcus abyssi*;Pfu = *Pyrococcus furiosus*;Rn = *Rattus norvegicus*;Sc = *Saccharomyces cerevisiae*;Sp = *Schizosaccharomyces pombe*;Ss = *Sulfolobus solfataricus*;Tv = *Thermoplasma volcanium.* The number in each node indicatates the confidence value of that branch after bootstrapping the phylogenetic tree.



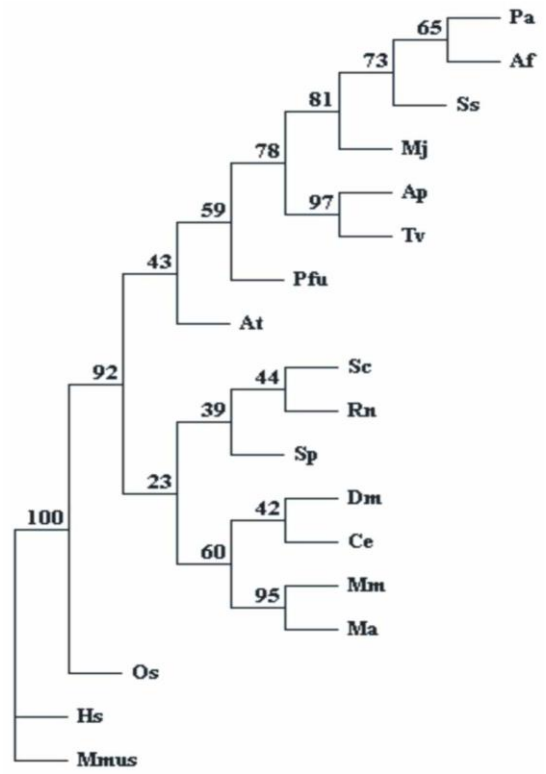
Supplementary Figure 8. Bootstrapped Phylogenetic tree of Uracil DNA glycosylase. Ap = *Aeropyrum pernix*;Af = *Archaeoglobus fulgidus*;At = *Arabidopsis thaliana* ;Ce = *Caenorhabditis elegans* ;Dm = *Drosophila melanogaster*;Hu = *Homo sapiens*;Ma = *Methanosarcina acetovorans*;Mj = *Methanocaldococcus jannaschii*;Mm = *Methanosarcina mazei*;Mmus = *Mus musculus*;Os = *Oryza sativa*;Pa = *Pyrococcus abyssi*;Pfu = *Pyrococcus furiosus*;Rn = *Rattus norvegicus*;Sc = *Saccharomyces cerevisiae*;Sp = *Schizosaccharomyces pombe*;Ss = *Sulfolobus solfataricus*;Tv = *Thermoplasma volcanium.* The number in each node indicatates the confidence value of that branch after bootstrapping the phylogenetic tree.



Supplementary Figure 9. Bootstrapped Phylogenetic tree of XPG endonuclease. Ap = *Aeropyrum pernix*;Af = *Archaeoglobus fulgidus*;At = *Arabidopsis thaliana* ;Ce = *Caenorhabditis elegans* ;Dm = *Drosophila melanogaster*;Hu = *Homo sapiens*;Ma = *Methanosarcina acetovorans*;Mj = *Methanocaldococcus jannaschii*;Mm = *Methanosarcina mazei*;Mmus = *Mus musculus*;Os = *Oryza sativa*;Pa = *Pyrococcus abyssi*;Pfu = *Pyrococcus furiosus*;Rn = *Rattus norvegicus*;Sc = *Saccharomyces cerevisiae*;Sp = *Schizosaccharomyces pombe*;Ss = *Sulfolobus solfataricus*;Tv = *Thermoplasma volcanium.* The number in each node indicatates the confidence value of that branch after bootstrapping the phylogenetic tree.



Supplementary Figure10. Bootstrapped Phylogenetic tree of RFC3. Ap = *Aeropyrum pernix*;Af = *Archaeoglobus fulgidus*;At = *Arabidopsis thaliana* ;Ce = *Caenorhabditis elegans* ;Dm = *Drosophila melanogaster*;Hu = *Homo sapiens*;Ma = *Methanosarcina acetovorans*;Mj = *Methanocaldococcus jannaschii*;Mm = *Methanosarcina mazei*;Mmus = *Mus musculus*;Os = *Oryza sativa*;Pa = *Pyrococcus abyssi*;Pfu = *Pyrococcus furiosus*;Rn = *Rattus norvegicus*;Sc = *Saccharomyces cerevisiae*;Sp = *Schizosaccharomyces pombe*;Ss = *Sulfolobus solfataricus*;Tv = *Thermoplasma volcanium.* The number in each node indicatates the confidence value of that branch after bootstrapping the phylogenetic tree.



Supplementary Figure 11. Bootstrapped Phylogenetic tree of WRN helicase. Ap = *Aeropyrum pernix*;Af = *Archaeoglobus fulgidus*;At = *Arabidopsis thaliana* ;Ce = *Caenorhabditis elegans* ;Dm = *Drosophila melanogaster*;Hu = *Homo sapiens*;Ma = *Methanosarcina acetovorans*;Mj = *Methanocaldococcus jannaschii*;Mm = *Methanosarcina mazei*;Mmus = *Mus musculus*;Os = *Oryza sativa*;Pa = *Pyrococcus abyssi*;Pfu = *Pyrococcus furiosus*;Rn = *Rattus norvegicus*;Sc = *Saccharomyces cerevisiae*;Sp = *Schizosaccharomyces pombe*;Ss = *Sulfolobus solfataricus*;Tv = *Thermoplasma volcanium.* The number in each node indicatates the confidence value of that branch after bootstrapping the phylogenetic tree.