

Towards an understanding of protein-protein interaction network hierarchies. Analysis of DnaN (β)-binding peptide motifs in members of protein families interacting with the eubacterial processivity clamp, the β subunit of DNA Polymerase III

Brian P. Dalrymple, Gene Wijffels, Kritaya Kongsuwan and Phil Jennings¹

CSIRO Livestock Industries, 120 Meiers Road, Indooroopilly, QLD 4068, Australia

¹Now at Peptech Ltd, 1/35-41 Waterloo Road, North Ryde, NSW 1670, Australia

Brian.Dalrymple@csiro.au

Abstract

The consensus pentapeptide QL[SD]LF is a major component in the interaction of a number of families of proteins with the eubacterial DNA-clamp protein, DnaN (the β -subunit of DNA Polymerase III holoenzyme). Rankings of the motifs were established using the program MEME. The distribution of ranking of motifs in the PolC, DinB2 and UmuC protein families were shown to be significantly skewed to higher rankings (ie closer to the consensus), whilst motifs in the DinB1 and DnaE2 protein families were shown to be significantly skewed to lower rankings. Sub-division of the families on the basis of domain architecture, clustering of sequences, presence or absence of members of other DnaN-binding protein families and phylogenetic positions, has identified a number of relationships. Species containing PolC have DnaE1 and DinB1.1 proteins with significantly lower ranking motifs and MutS1 proteins with significantly higher-ranking motifs than average. These observations suggest that this group of eubacteria have a different hierarchy of interactions of DnaN-binding proteins from that in the majority of other species. Members of a new family of proteins of unknown function, Duf72, also have high-ranking putative DnaN-binding motifs. Motifs in a second new family containing putative DnaN-binding sites, DinB3, rank at the lower end.

Keywords: protein-protein interaction, DNA Polymerase III, sliding clamp.

Introduction

The replication of DNA in eubacteria involves many proteins organised in a complex multifunctional machine known as the replisome. The central enzyme is the DNA Polymerase III holoenzyme. In *Escherichia coli* this enzyme contains 10 different subunits, whilst in most other bacteria only seven subunits have been identified. ¹In *E. coli*, and probably in most other eubacteria, the DnaE orthologue (α subunit) is the main replicative

polymerase, but in many gram positive organisms a distinct, but related enzyme, PolC is proposed to be the main replicative enzyme replacing DnaE (Bruck and O'Donnell, 2000).

The processivity of the replisome is conferred by the DnaN (β) subunit of DNA Polymerase III, which forms a clamp around the DNA (Kong et al., 1992). DnaE and PolC from a number of different species have been shown to bind to DnaN (Low et al., 1976; Bruck and O'Donnell, 2000; Klemperer et al., 2000; Bruck et al., 2002; Bullard et al., 2002; Noiro-Gros et al., 2002), suggesting that this will be a universally conserved interaction.

DnaN is loaded onto DNA by a clamp loader comprising single subunits of HolA (δ subunit of DNA Polymerase III) and HolB (δ' subunit of DNA Polymerase III) and four subunits of DnaX (τ/γ subunits of DNA Polymerase III) (Bruck and O'Donnell, 2000). In the clamp loader the HolA subunit binds to DnaN (Jeruzalmi et al., 2001). The interaction of HolA with DnaN has been demonstrated in a number of different species (Bruck and O'Donnell, 2000; Klemperer et al., 2000; Bruck et al., 2002; Bullard et al., 2002; Noiro-Gros et al., 2002), suggesting that this will also be a universally conserved interaction.

A number of other components of the DNA replication and repair complexes of *E. coli* have been reported to interact with DnaN; the DNA repair polymerases PolA (Lopez de Saro and O'Donnell, 2001), PolB (Hughes et al., 1991; Bonner et al. 1992), UmuD₂UmuC (Sutton et al., 1999; Tang et al., 2000) and DinB (Tang et al., 2000, Wagner et al., 2000), the regulator of initiation of replication, IdaB/Hda (Kato and Katayama, 2001), the mismatch recognition protein MutS1 (Lopez De Saro and O'Donnell, 2001) and the DNA ligase, LigA (Lopez De Saro and O'Donnell, 2001). In *Bacillus subtilis* YqjH (a member of the DinB family), and proteins from two additional families, MutL and YabA, have also been shown to bind to DnaN (Noiro-Gros et al., 2002). MutL, like MutS, is involved in the mismatch repair pathway (Yang et al., 2000). YabA also binds to DnaA and, although unrelated at the sequence level to IdaB/Hda, has been proposed to also regulate the initiation of rounds of replication (Noiro-Gros et al., 2002)

The binding of the *E. coli* DnaE, PolB, UmuC, DinB and UmuC proteins to DnaN has been shown to involve a pentapeptide motif with the consensus sequence QL[SD]LF (Dalrymple et al., 2001, Lenne-Samuel et al.,

2002). This motif has also been identified in members of the PolC protein family, but not in the UmuD protein family (Dalrymple et al 2001). It is probable that UmuD interacts with DnaN via a completely different mechanism (Sutton et al., 2002). The motif(s) involved in the binding of members of the PolA, LigA, MutL, YabA and IdaB/Hda families of proteins to DnaN have not been investigated. A related, but distinct, tripeptide motif (SLF), has been identified in members of the HoloA family of clamp loading subunits (Dalrymple et al., 2001).

The motif appears to be conserved across the complete range of eubacterial evolution. Mix and match experiments with DnaE, PolC and DnaN proteins from a range of species (Bruck and O'Donnell, 2000; Klemperer et al., 2000; Kongsuwan et al., unpublished) show that proteins from diverse species can interact. These results suggest that the motif, and its binding site on DnaN, are a universal system and that each protein family, if not the complete data set can be treated as a single data set.

The closer a peptide is to the consensus sequence the stronger the binding to DnaN (Dalrymple et al., 2001; Wijffels et al., unpublished). This suggests that proteins containing motifs with good matches to the consensus sequence may generally bind more strongly to DnaN than those with poor matches. The identified motifs vary widely in similarity to the consensus, with many proteins containing motifs that have little, or no binding activity in *in vitro* peptide assays (Dalrymple et al., 2001).

To increase our understanding of the interactions of proteins with DnaN we have undertaken a detailed bioinformatics analysis of DnaN-binding proteins.

Results

1.1 Definition of protein families

The growing number of eubacterial genome sequences now provides access to large data sets of orthologous proteins. The data sets used in this work were compiled from exhaustive searches of the completed and in progress microbial genome databases at the NCBI, TIGR, the Sanger Center and the DOE Joint Genome Institute. The BLAST searches were initiated with the founding members of each of the families (Table 1). The gene families were further grouped into superfamilies on the basis of their amino acid alignments and clustering using CLUSTAL. The DNA polymerase superfamilies are defined in Burgers et al., 2001.

super family	family	founder	species	motif location
C	DnaE1	DnaE	<i>E. coli</i>	central
	DnaE2	DnaE2	<i>Mycobacterium tuberculosis</i>	central
B	PolC	PolC	<i>B. subtilis</i>	C-term ¹ .
	PolB2	PolB	<i>E. coli</i>	C-term.
Y	DinB1	DinB	<i>E. coli</i>	close to C-term.
	DinB2	YqjW	<i>B. subtilis</i>	close to C-term.

UmuC	UmuC	<i>E. coli</i>	close to C-term.
MutS	MutS1	MutS	<i>E. coli</i>
			at or close to C-term.

¹carboxy-terminus of the protein

Table 1: Protein families previously identified as containing DnaN-binding peptide motifs

1.2 Sequence and putative structural context of the DnaN-binding motifs

From the alignments of the amino acid sequences of the members of each of the protein families the region of the global alignment containing the putative DnaN-binding motif was identified as previously described (Dalrymple et al., 2001). The locations of the DnaN-binding motifs are listed in table 1.

The putative DnaN-binding peptides are located very close to, or at, the carboxy-terminus of members of several families of proteins. Many eukaryotic proteins, and some bacteriophage DNA polymerases, contain an equivalent clamp protein binding motif in a similar location (Warbrick et al., 2000). By analogy with the structures of a number of these complexes (Hosfield et al., 1998; Shamoo and Steitz, 1999) it is predicted that the DnaN-binding peptide at the carboxy-termini of proteins will be relatively unstructured in the absence of binding. In these proteins the motifs may represent the major region of contact between the DnaN-binding protein and DnaN itself.

In members of the DnaE1 and DnaE2 families the peptide is flanked by conserved domains. It is likely that the DnaN-binding peptides in these proteins are more constrained than the carboxy-terminally located DnaN-binding peptides.

In most members of the MutS1 protein family the putative DnaN-binding peptides are located in a region of highly variable sequence and length. This region is located between the catalytic domain and a short conserved domain of approximately 25 amino acids at the carboxy-terminus of the proteins. Thus these DnaN-binding peptides are also likely to be in a relatively unconstrained environment.

1.3 MEME analysis of putative DnaN-binding motifs

Twenty five amino acid long sequences containing the putative DnaN-binding motif and ten flanking residues from either side for all members of each of the protein families described in table 1 were then analysed. The MEME (Bailey and Elkan, 1994) motif identification program (version 3) was run with the following parameters, zero or one motif per sequence and motif 2-25 amino acids wide for the first analysis. The maximum size of the motif was then reduced until no further members of the protein family were included in the list of significant matches. The first line of consensus sequence

from the MEME search was taken as the core motif for each of the protein families (Table 2).

family	size	hits	motif
DnaE1	86	72	<u>GQxSLFG</u>
DnaE2	19	19	<u>QLPLFADAPAIE</u>
PolC	22	20	GCLGDLPEDN <u>QLSLF</u>
PolB2	15	14	EDDFATLLTG <u>QLGLF</u>
DinB1	70	57	<u>RQLSLF</u>
DinB2	11	11	LSNLIIDSESE <u>QLSLF</u> FDDx EER
UmuC	31	27	<u>QLNLFDEVAPRPGSE</u>
MutS1	67	58	<u>QLSLF</u>

Table 2: Peptide motifs identified using MEME

The predicted peptide motifs from each of the families of proteins contain a clearly similar pentapeptide motif, underlined in table 2. This motif conforms to the previously identified consensus sequence of QL[SD]LF for DnaN-binding peptide motifs (Dalrymple et al., 2001) and confirms the validity of analysing these motifs as a functionally related group of peptides.

However, one difference is readily apparent on visual inspection of the sequences. The DnaE1 motif is the only one that does not have a conserved leucine residue identified at position 2 of the pentapeptide motif. Analysis of the frequency of leucine at position 2 of the pentapeptide in DnaE1 sequences v. the sequences of the motifs from the other protein families confirmed that this difference was highly significant with a z score of 7.5 in a test of proportions for DnaE1 motifs v the rest. However, the biological significance of this is not clear. In *in vitro* assays peptides containing leucine at position 2 bind more strongly to DnaN than those with other amino acids (Dalrymple et al., 2001, Wijffels et al., unpublished).

1.4 Some protein families have significantly biased distributions of peptide motif ranks

The complete data set of 25 amino acid fragments was then analysed using MEME and the peptides ranked on the basis of their MEME probability scores. These scores provide a measure of the similarity of a particular motif to the position specific substitution matrix (PSSM) of the complete data set. The higher the rank the closer a motif is to the MEME consensus based on the PSSM. The distribution of ranks for the motif from each protein family was then compared to the distribution of ranks of the motifs from all of the other protein families. P-values were calculated for the Kruskal-Wallis rank sum test using the R statistical package. For significant p-values (at the 99% confidence interval), highlighted in bold, the relationship of the family ranks to the ranks in the complete data set was also recorded (Table 3). Families with high ranks have a distribution of ranks of motifs skewed towards the consensus sequence.

family	p-value	ranks
DnaE1	0.09	-

DnaE2	0.0012	low
PolC	2.5e-06	high
PolB2	0.012	-
DinB1	4.3e-06	low
DinB2	0.0046	high
UmuC	0.007	high
MutS1	0.36	-

Table 3: p-values for Kruskal-Wallis test for the named family v. the complete data set of peptide motifs not including the named protein family

The motifs present in the PolC family exhibited a very significantly higher ranking than the combined data set of all of the other families. The PolC motif is located at, or very close to, the carboxy-terminus of the proteins. To test the hypothesis that location, rather than protein family might be responsible for the observed distribution of rankings the data was divided on the basis of location of motif in the protein. No significant difference in distribution of rankings was observed between motifs with five or less amino acids to the end of protein and motifs further away from the ends (data not shown).

In contrast the motifs in the members of the DinB1 family exhibited a very significantly lower ranking than the combined data set of all of the other families. The data from the members of the DinB1 family is discussed in more detail in section 1.8.

1.5 Species containing both PolC and DnaE1 have DnaE1 proteins with significantly low ranking DnaN-binding peptide motifs

The initial grouping of the proteins into families was based on the assumption that all members of the same sequence clusters were likely to be functional orthologues. Whilst, this is unlikely to be the case for the DinB1 family, given the presence of multiple proteins in single species and different domain architectures, it is not obvious from the sequence that not all of the DnaE1 proteins may be true functional orthologues. However, it has recently been shown that although all species of bacteria contain a member of the DnaE1 family they may not be functional orthologues (Koonin and Bork., 1996; Bruck and O'Donnell, 2000; Dervyn et al., 2001; Inoue et al., 2001).

To investigate in more detail possible interactions between the presence and absence of representatives of particular protein families an analysis of the MEME rankings of p-values for the DnaE1 data set alone was carried out. The rankings distributions of the DnaE1 data divided on the basis of the presence of a member or members of each of the other protein families were then compared (+PolC species, Table 4). For significant p-values (at the 99% confidence interval) the relationship of the family ranks to the ranks in the complete data set was recorded (Table 4). In a second analysis (-PolC species), to remove the contributions due to the low ranking values in species containing PolC, the DnaE1 motif scores from

species containing PolC proteins were not included in the calculations (Table 4).

Family	p-value DnaE1 motif ranks		Ranks
	+PolC species	-PolC species	
DnaE2	0.0059	0.13	high
PolC	4.6e-08	na	low
PolB2	0.023	0.22	-
DinB1	0.069	1.1e-05	high
DinB2	1.6e-06	na	low
UmuC	0.68	0.42	-
MutS1	0.96	0.97	-

Table 4: p-values for Kruskal-Wallis test - DnaE1 motifs, including (+PolC) and excluding (-PolC) motifs from PolC containing species

In species containing PolC and DinB2, the putative DnaN-binding motifs in the DnaE1 proteins exhibited a distribution skewed towards low rankings. Since all species containing DinB2 also contained PolC these results are not independent and the presence of either or both of these proteins could be associated with the low rankings of the DnaE1 motifs, ie. poor matches to the MEME consensus.

One possible explanation for the high ranking PolC motifs and the low ranking DnaE1 motifs in PolC containing species is provided by the proposed different roles of PolC and DnaE1 in *B. subtilis*, *Staphylococcus aureus* and *Streptococcus pyogenes*. PolC is proposed to undertake leading strand synthesis, which requires a very stable interaction between the polymerase subunit and DnaN. DnaE1 is proposed to undertake the lagging strand synthesis (Bruck and O'Donnell, 2000; Dervyn et al., 2001; Inoue et al., 2001), which requires much more frequent disassociation of the polymerase subunit and DnaN. In species not containing PolC the DnaN-binding might be expected to be a compromise between the two competing requirements of leading and lagging strand synthesis. The interaction of DnaE1 and DnaN from *S. pyogenes* has been demonstrated with an increase in processivity of the polymerase (Bruck and O'Donnell, 2000). Thus, although the proposed DnaN-binding motif in *S. pyogenes* (LGSLF) has a low ranking, it is presumably still sufficient to mediate interaction with DnaN.

MEME analysis of the DnaE1 motifs from species divided on the basis of the occurrence of PolC (DnaE1+pc), or not (DnaE1-pc), was carried out (Table 5).

Family	size	hits	motif
DnaE1-pc ¹	68	68	QxAKAEASGQxDLFGGLxDxxE
DnaE1+pc	18	11	FVEEDGSLFD
DnaE1+pc+LF ²	12	12	FVEEDGSLFD
DnaE1+pc-LF	6	-	-

¹containing PolC (+pc) or not containing PolC (-pc)

²containing LF (+LF) or not containing LF (-LF)

Table 5: Peptide motifs in the subfamilies of the DnaE1 protein family

However, whilst all of the motifs from the PolC negative species formed a cluster this was not the case from PolC positive strains. The DnaE1 sequences from PolC positive strains were then further divided on the basis of their p-values and conservation of the leucine and phenylalanine at positions four and five and submitted to MEME (Table 5). No significant motif could be identified in the other sequences.

The distribution of species containing the motifs assigned to the two +pc clusters was examined (Table 6).

Cluster	species	
+LF	<i>Thermotoga maritima</i>	
	<i>Clostridium difficile</i>	
	<i>Bacillus halodurans</i>	
	<i>Bacillus subtilis</i>	
	<i>Staphylococcus aureus</i>	
	<i>Staphylococcus epidermidis</i>	
	<i>Lactococcus lactis</i>	
	<i>Streptococcus pyogenes</i>	
	<i>Streptococcus pneumoniae</i>	
	<i>Mycoplasma genitalium</i>	
	<i>Mycoplasma pneumoniae</i>	
	<i>Clostridium acetobutylicum</i>	
	-LF	<i>Carboxydotherrmus hydrogenoformans</i>
		<i>Bacillus stearothermophilus</i>
<i>Bacillus anthracis</i>		
<i>Enterococcus faecalis</i>		
<i>Enterococcus faecium</i>		
	<i>Ureaplasma urealyticum</i>	

Table 6: Distribution of species containing PolC on the basis of the type of motif contained in DnaE1.

There was no readily identifiable relationship between the clusters and the phylogenetic positions of the species. The data set is probably too small to analyse other possible relationships.

1.6 An association between DnaE2 and DinB1 and the DnaE1 ranking?

A significant result was observed for DnaE2 in the whole data set, but not after the removal of the DnaE1 motifs from PolC containing species (Table 4). This probably reflects the influence of the low ranking motifs from PolC contain species, none of which contain DnaE2. Thus this result is not considered of biological significance.

Another highly significant result is that after removal of the DnaE1 motifs from PolC containing species the rankings in species containing members of the DinB1 family are higher than in species not containing a member of the DinB1 family. (Table 4). This suggests that there may be an interaction between DinB1 family proteins and DnaE1 family proteins in species that do not contain PolC, most of which do contain a member of the DinB1 protein family.

1.7 Detailed analysis of the MutS1 family

The members of the MutS1 family exhibited a number of different domain organisations with a number of species, *Thermus* sp. and *Buchnera* sp. apparently not containing a DnaN-binding region. However, unlike the DinB1 family (see section 1.8), only one gene encoding a member of the MutS1 family has been identified in any single genome of a eubacterium. To break up the MutS1 family into subfamilies of possibly different functionally orthologous proteins, protein groups were created on the basis of domain architecture, two subfamilies of the MutS1 family containing putative DnaN-binding sites were identified.

- MutS1.1, no carboxy-terminal extension, a small number of species.
- MutS1.2, carboxy-terminal extension, the majority of species.

As described above, the 25 amino acid regions containing the putative pentapeptide, or the equivalent regions for each member of the MutS1 protein family, were submitted to MEME grouped into the subfamilies and analysed as previously described (Table 7).

family/subfamily	size	hits	motif
MutS1	67	57	<u>QLSLF</u>
MutS1.1	4	4	<u>QLDLF</u>
MutS1.2	63	54	<u>QLSLF</u>
MutS1.2+mc ¹	54	54	Px <u>QLSLFAAAP</u> xPE
MutS1.2-mc	9	-	-

¹in MEME consensus (+mc) not in consensus (-mc)

Table 7: Peptide motifs in subfamilies of the MutS1 protein family

Comparison of the rankings distributions of the MEME probability scores from the MutS1.1 and MutS1.2 subfamilies (using the complete MutS1 family data set) using the Kruskal-Wallis test produced a p-value of 0.41, indicating that there was no significant difference. Although MutS1.1 contains only a small number of

sequences it appears that the presence or absence of the short conserved carboxy-terminal domain, of currently unknown function, has no direct impact on the nature of the DnaN-binding motif.

Inspection of the MEME results shows that there are a number of very low scoring motifs in the MutS1 subfamily. The MEME analysis was repeated separating the motifs with significant (MutS1+mc) MEME p-values from those without (MutS1-mc). No consensus sequence was identified in the MutS1-mc group (Table 7). The species containing these motifs do not form a monophylogenetic cluster (Table 8).

Cluster	species
MutS1.2-mc	<i>Aquifex aeolicus</i>
	<i>Aquifex pyrophilus</i>
	<i>Borrelia burgdorferi</i>
	<i>Lengionella pneumophiia</i>
	<i>Shewanella putrefaciens</i>
	<i>Pasteurella multocida</i>
	<i>Actinobacillus actinomycetemcomitans</i>
	<i>Clostridium difficile</i>
	<i>Clostridium acetobutylicum</i>

Table 8: Species in the MutS1.1-mc subfamily

Possible relationships between the presence of members of other protein families of DnaN-binding proteins and different rankings of motifs were then investigated. The MutS1 data set was divided up on the basis of the presence or absence of members of the other families and the rankings distribution compared using the Kruskal-Wallis test (Table 9). Results significant at the 99% confidence interval are highlighted in bold.

Family	p-value	MutS1 ranks	ranks
DnaE2	0.37		-
PolC	0.0021		high
PolB2	0.079		-
DinB1	0.72		-
DinB2	0.0013		high
UmuC	0.56		-

Table 9: p-values for Kruskal-Wallis test - MutS1 motifs

The only significant difference was in the presence, or absence, of PolC/DinB2, which have overlapping distributions. Species containing these proteins had a distribution of rankings skewed to the higher end compared to the rest of the data set. The reason for such a relationship is not clear as DinB1 and MutS1 are involved in different pathways in the cells.

1.8 Detailed analysis of the DinB1 family

The DinB1 family also contained a significant number of members that did not have significant matches to the

motif sequence identified using MEME. Inspection of the family showed that it contained members with several different domain organisations. To break up the DinB1 family into subfamilies of probable functionally orthologous proteins, protein groups were created on the basis of amino acid sequence clustering using CLUSTAL (using only the common core region of the proteins) and domain architecture. Five subfamilies of the DinB1 family were identified.

- DinB1.1, no amino-terminal or carboxy-terminal extensions.
- DinB1.2, carboxy-terminal extension type A.
- DinB1.3, carboxy-terminal extension type A and cysteine-rich amino-terminal domain.
- DinB1.4, carboxy-terminal extension type B
- DinB1.5, apparently no DnaN-binding region

The type A and type B carboxy-terminal domains in the DinB1 subfamilies have unrelated amino acid sequences. However, the type A domain is related to the equivalent regions of members of the DinB2 and UmuC families of proteins and thus probably represents the ancestral organisation. Representatives of the DinB1.1 and DinB1.2 subfamilies have been shown to bind to DnaN (Tang et al., 2000; Wagner et al., 2000; Lenne-Samuel et al., 2002; Noiro-Gros et al., 2002), no experiments have been described for the members of the other subfamilies. The distribution of the DinB1.1 and DinB1.2 subfamilies is interesting. The DinB1.2 containing species are sporadically distributed through the DinB1.1 species, but with one exception the DinB1.2 sequence clusters do not contain DinB1.1 sequences. Indeed, *Bacillus* sp. (contain members of the DinB1.2 family) and *Streptococcus* sp. (contain members of the DinB1.1 family) are closely related, but the DinB proteins do not cluster. Sequence clustering appears to be a function of domain architecture, rather than the phylogenetic position of the species containing the protein. This result suggests horizontal gene transfer, and/or an ancient duplication generating two genes, each encoding protein of potentially different functions, followed by gene loss. With one exception no species has genes encoding both a DinB1.1 and DinB1.2 family protein. In contrast, *Mycobacterium* sp. contain members of the DinB1.2, DinB1.4 and DinB1.5 families and a number of species containing members of the DinB1.3 family also contain members of the DinB1.1 family.

As described above the 25 amino acid regions containing the putative pentapeptide, or the equivalent regions, for each member of the DinB1 protein family were submitted to MEME grouped into the subfamilies. In this case all groups were analysed for motifs 2-5 amino acids long. The results were not as clear cut as in the analysis of other families, for instance initially no pentapeptide motifs related to QL[SD]LF was identified in the DinB1.3 and DinB1.4 families. The analysis was repeated with relaxed length parameters for these two subfamilies (Table 10).

family/ subfamily	size	hits	motif	p-value ¹
DinB1	70	57	RQLSLF	
DinB1.1	35	26	QLELEF	0.34
DinB1.2	24	24	QLDLF	0.15
DinB1.3	7	6	DLLDPQ	2.2e-5
DinB1.4	4	4	VGFSGLSDIRQESLFPD	0.6

¹Kruskal-Wallis p-values for the comparison of distribution of ranks for each subfamily v. the rest, using the complete DinB1 data set.

Table 10: Peptide motifs in subfamilies of the DinB1 protein family

The motif identified in the DinB1.3 subfamily does not appear to be related to the QL[SD]LF motif. Thus members of the DinB1.3 subfamily may not bind to DnaN, or bind via a new DnaN-binding motif. Experimental testing of the interaction of members of the DinB1.3 family with DnaN is required.

In the DinB1.1, DinB1.2 and DinB1.4 subfamilies the presence, or absence, of different amino- and carboxy-terminal domains does not lead to differences in the distributions of motif ranks.

Visual inspection of the output for the DinB1.1 subfamily identified a correlation between low scores and species that have been shown to contain DinB2 and/or PolC proteins. The DinB1.1 motifs were then divided on the basis of the presence of DinB2 and reanalysed with MEME (Table 11).

subfamily	size	hits	motif
DinB1.1-DinB2	27	27	×EQLELEF
DinB1.1+DinB2	8	8	VTALEDSVLK

Table 11: Peptide motifs in the DinB1.1 subfamily, divided on the presence or absence of DinB2

In species that do not contain DinB2, a motif consensus sequence similar to the global consensus sequence was identified. In species that do contain DinB2, a motif with no apparent similarity to the global consensus motif was identified. Thus, members of the DinB1.1 subfamily in species containing DinB2 may not bind to DnaN, or if they do it is via a different motif. It is tempting to speculate that since the new motif is located in the equivalent region of DinB1.1 that it might be a new DnaN-binding motif. Experimental testing of the interaction of members of the DinB1.1 family from DinB2 positive species with DnaN is required.

Since the distribution of members of the DinB2 and PolC protein families are almost the same, a Kruskal-Wallis test was conducted on the rankings of the MEME p-values for the data from the DinB1.1 subfamily. To compare the DinB1.1 and DinB1.2 subfamilies an independently analysis was undertaken on the DinB1.2 subfamily (Table 12).

subfamily	comparison	p-value	Ranks
DinB1.1	+/-DinB2	0.0041	+ low
	+/-PolC	0.0031	+ low
DinB1.2	+/-DinB2	0.074	-
	+/-PolC	0.56	-

Table 12: p-values in the Kruskal-Wallis test

No significant difference in distribution of ranks was observed for the DinB1.2 subfamily with either DinB2 or PolC. In contrast, for the DinB1.1 subfamily a significant difference was observed with both DinB2 and PolC.

1.9 Identification of a putative DnaN-binding motif in the DinB3 protein family

During the analysis of the members of the Y superfamily of DNA polymerases an additional protein family was identified in the eubacteria. This family, founder Rv3394c from *M. tuberculosis*, has been designated the DinB3 family. Although the function of members of the DinB3 family has to our knowledge not been investigated the amino acid sequence similarity suggests that it is a DNA polymerase. Interestingly, the gene encoding DinB3 is almost always located adjacent to a gene encoding a member of the DnaE2 family, suggesting that these two proteins are involved in the same DNA replication or repair pathway. The regions of the members of the DinB3 family equivalent to the regions of the DinB1, DinB2 families shown to containing DnaN-binding motifs were analysed using MEME (Table 13). However, less than half of the sequences were included in the consensus. The included (+mc) and not included (-mc) sequences were then analysed separately (Table 13).

Family/subfamily	size	hits	motif
DinB3	21	9	LFDEP
DinB3+mc ¹	9	9	FVPQHDEL LFDEP
DinB3-mc	12	12	VEPLAAAQ LFDEP DG

¹in MEME consensus (+mc) not in consensus (-mc)

Table 13: Peptide motifs in the DinB3 protein family and subfamilies.

The MEME analysis identified two motifs apparently representing the first or the second halves of the QL[SD]LF motif. This result suggests that it is quite likely that the members of the DinB3 family do bind to DnaN. This prediction awaits experimental testing.

The distribution of species between the two clusters is shown below (Table 14).

Cluster	species
DinB3+mc	<i>Bordetella bronchiseptica</i>
	<i>Burkholderia cepacia</i>
	<i>Burkholderia mallei</i>
	<i>Ralstonia metallidurans</i>
	<i>Methylococcus capsulatus</i>
	<i>Pseudomonas aeruginosa</i>
	<i>Pseudomonas putida</i>
	<i>Pseudomonas syringae</i>
	<i>Pseudomonas fluorescens</i>
	DinB3-mc
<i>Rhodopseudomonas palustris</i>	
<i>Mesorhizobium loti</i>	
<i>Brucella suis</i>	
<i>Sinorhizobium meliloti</i>	
<i>Caulobacter crescentus</i>	
<i>Rhodobacter capsulatus</i>	
<i>Sphingomonas aromaticivorans</i>	
<i>Acidithiobacillus ferrooxidans</i>	
<i>Mycobacterium smegmatis</i>	
<i>Mycobacterium tuberculosis</i>	
<i>Corynebacterium diphtheriae</i>	

Table 14: Distribution of species between the two motif-based subfamilies of the DinB3 family

The clustering of the two different motifs is along phylogenetic lines with the DinB3 motifs containing LF forming a cluster within the more deeply branched distribution of the other motif.

1.10 Searching for putative QL[SD]LF-type DnaN-binding motifs in the DnaE3 family

The searches of the databases for members of the DnaE1 family also identified a further new family in addition to the DnaE2 family. The members of the DnaE3 family cluster separately from the DnaE1 and DnaE2 families. The members include a plasmid encoded protein from *Yersinia pestis* and a bacteriophage encoded enzyme from *B. subtilis* (YorL, the founding member). MEME did not identify any significant motifs from the regions of these proteins equivalent to the region containing the DnaN-binding motifs from members of the DnaE1 family.

1.11 Characterisation of the SLF-type DnaN-binding motif in the HoloA family

The HoloA subunit of the clamp (DnaN)-loader contacts the DnaN protein, with the major site of interaction, an LF motif being identified by bioinformatics analysis (Dalrymple et al., 2001) and subsequently from structural

studies (Jeruzalmi et al., 2001). MEME analysis of the HolA alignments identified a conserved motif (Table 15).

Family	size	hits	motif
HolA	84	81	LLEELxSxSLFADRRLLVVLR

Table 15: Peptide motifs in the HolA protein family identified using MEME

Unlike most of the other protein families containing DnaN-binding motifs this region is in a folded structural domain of the protein. The interaction also involves important interactions outside of the conserved SLF motif (Jeruzalmi et al., 2001). The interaction of HolA and DnaN requires a change in the conformation of both proteins. The conformational change in DnaN leads to separation of the subunits of the dimer at one interface, required for loading DnaN onto DNA (Jeruzalmi et al., 2001). It would seem unlikely that the interaction of DnaE1, PolC etc. with DnaN would lead to a similar change, as processivity of DNA synthesis presumably requires a stable DnaN dimer.

1.12 Searching for putative QL[SD]LF-type DnaN-binding motifs in other families

The MutL, PolA and LigA protein families are widely distributed in the eubacteria. The YabA protein family is confined to a small group of bacteria related to *B. subtilis*. The DnaA2 protein family is widely distributed in the Proteobacteria. The founder proteins are described in Table 16.

family	founder	function	species
DnaA2	IdaB/Had	Regulation of initiation of DNA replication	<i>E. coli</i>
	/YfgE		
YabA	YabA	Regulation of initiation of DNA replication?	<i>B. subtilis</i>
PolA	PolA	DNA Polymerase I	<i>E. coli</i>
LigA	Lig	DNA ligase	<i>E. coli</i>
MutL	MutL	Mismatch repair pathway	<i>E. coli</i>

Table 16: Founders of additional families of DnaN-binding protein families

The amino acid sequences of members of the families of proteins were aligned using CLUSTAL. The resulting alignments were then searched for conserved QL[SD]LF motifs. Although isolated examples of the consensus, or related sequences, were identified in many members of these families no unambiguous putative DnaN-binding pentapeptide motif was identified in any of the families.

1.13 Identification of a new family of eubacterial proteins containing a conserved QL[SD]LF motif, that may be involved in binding to DnaN

Database searches looking for additional families of proteins potentially containing a QL[SD]LF motif were carried out using the consensus sequence and related sequences. The Pattenprot server (http://pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_patinprot.html) and the BLAST search for short nearly exact matches at the NCBI (<http://www.ncbi.nlm.nih.gov>) were used. This

approach identified a number of new candidate protein families. In one of these, the Duf72 protein family (Pfam pf01904), the matches appear likely to have identified a new family containing DnaN-binding proteins. The motif is located at the carboxy-terminus of the proteins, but unlike the other families of proteins is only present in a small number of the members. In particular, in two clusters of species closely related to *E. coli* and to *B. subtilis*. MEME analysis of the region from proteins containing a carboxy-terminal extension identified a motif (IEYxGLAPEQLDLF) with a very good match to the overall consensus.

The Duf72 protein family is diverse, with multiple copies present in some genomes, suggesting that not all members are true orthologues. The two clusters containing putative DnaN-binding sites are distinct, but each group is tightly clustered within the CLUSTAL trees, excluding any protein sequences without the motif.

Despite extensive analysis using NCBI PSI BLAST no other families of proteins related to Duf72 were identified. Thus the function of members of the Duf72 family remains unknown. The presence of a good DnaN-binding site in some members suggests that Duf72 may be involved in a DNA repair pathway of some description, but that interaction with between Duf72 and DnaN is not essential for function in most species of bacteria. The location of the motif at the carboxy-terminus suggests that it may be the major interaction site for the proposed interaction of Duf72 family members with DnaN (see section 1.2). Experimental testing of the interaction of members of the Duf72 family with DnaN is required.

1.14 A hierarchy of putative DnaN binding sites

The complete set of motifs and putative motifs from all of the protein families listed below (Table 17), a total of 453 sequences, were submitted for MEME analysis as described above. The families and subfamilies are listed below in order of the median MEME p-value for the grouping, from best to worst matches.

family/ subfamily	function
PolC ¹	Leading strand DNA polymerase
PolB2	Repair DNA polymerase
DinB2	Repair DNA polymerase?
Duf72	Unknown function
UmuC	Repair DNA polymerase
DinB1.4	Repair polymerase?
DnaE1-pc	Leading and lagging strand DNA polymerase
MutS1	Mismatch repair
HolA	Loading DnaN clamp
DnaE1+pc	Lagging strand DNA polymerase
DnaE2	Unknown DNA polymerase?

DinB3	Unknown DNA polymerase?
<i>DinB1.1+</i> <i>DinB2</i>	<i>Repair DNA polymerase?</i>
<i>DnaE3</i>	<i>Unknown DNA polymerase?</i>
<i>DinB1.4</i>	<i>Repair DNA polymerase?</i>

1. Protein families with pentapeptide motifs are indicated in **bold**, those with motifs shorter than five amino acids are indicated in normal text and those with different or no motifs identified are indicated in *italics*.

Table 17: Hierarchy of motifs in subfamilies of DnaN-binding proteins, with possible functional assignments

Conclusions

As a result of the analyses described above a number of observations can be made

- There appear to be good correlations between protein families and the distribution of motif rankings, but the distributions overlap substantially. Knowing the ranking of a motif, or its sequence, does not identify the protein family to which the protein containing the motif belongs.
- There does not appear to be a correlation between location of motif and motif rankings across families.
- In general the presence or absence of additional domains does not appear to affect ranking distributions - suggesting that these domains do not have an involvement in DnaN-binding?
- Episome encoded proteins DinB2 and UmuC (with the exception of DnaE3) have high rankings.
- Two new families containing putative DnaN-binding proteins, DinB3 and Duf72 have been identified.
- Possible new DnaN-binding motifs have been identified in DinB1.3 and DinB1.1+DinB2 subfamilies (but these proteins are also possibly non-binders to DnaN).

A number of possible relationships influencing level of similarity to the global consensus sequence for DnaN-binding motifs have also been identified. In the list below the protein families containing the affected motif are indicated in bold;

- **DnaE1** +/- (PolC or DinB2)
- (**DnaE1** - (PolC or DinB2)) +/- DinB1
- **MutS1** +/- (PolC or DinB2)
- **DinB1.1** +/- (PolC or DinB2)

A common feature of a number of these possible relationships is that they involve PolC or DinB2 (the two proteins have an almost congruent distribution). Unlike the MutS1 and DinB1 protein families, the members of the PolC and DinB2 protein families have a narrow distribution confined predominantly to species related to *B. subtilis*. Thus, the effects noted here may also reflect particular attributes of a cluster of related species. As

discussed above PolC containing organisms are predicted to have a number of differences in their DNA replication systems from the majority of bacteria. The putative relationships identified may reflect subtle differences in the interactions of proteins with DnaN in these species. However, whilst interactions of proteins with DnaN provide an attractive explanation for the observed correlations, the common evolutionary history of the genes, or the influence of another as yet unidentified protein in species related to *B. subtilis* cannot be ruled out.

Clearly the full sequences of a large number of eubacterial genomes has provided us with the raw data for a range of analyses that were not feasible only a few years ago. New and more sophisticated analysis methods are required in this exciting area of research. The extensive data set for the DnaN-binding sites provides an excellent starting point to address the questions around the hierarchies of protein-protein interactions and will help to drive the development of the new analysis methods and tools.

We do not know what other sites on these proteins are involved in the interactions of proteins with DnaN, or how poor a motif can be before interaction with DnaN does not occur. Our analysis has identified a number of candidate proteins for examination of this, such as the non-LF DnaE1 proteins from PolC positive species.

The analysis described above provides us with a glimpse of what the future may hold in the analysis of protein-protein interaction sites and the ability of the analysis to illuminate our understanding of the biology of such interaction networks.

References

- BAILEY, T.L. and ELKAN C. (1994): Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, 28-36, AAAI Press, Menlo Park, California.
- BONNER, C.A., STUKENBERG, P.T., RAJAGOPALAN, M., ERITJA, R., O'DONNELL, M., MCENTEE, K., ECHOLS, H. and GOODMAN, M.F. (1992): Processive DNA synthesis by DNA polymerase II mediated by DNA polymerase III accessory proteins. *Journal of Biological Chemistry* **267**(16): 11431-8.
- BRUCK, I. and O'DONNELL, M. (2000): The DNA replication machine of a gram-positive organism. *Journal of Biological Chemistry* **275**(37): 28971-83.
- BRUCK, I., YUZHAKOV, A., YURIEVA, O., JERUZALMI, D., SKANGALIS, M., KURIYAN, J. and O'DONNELL, M. (2002): Analysis of a multicomponent thermostable DNA polymerase III replicase from an extreme thermophile. *Journal of Biological Chemistry* **277**(19): 17334-48.
- BULLARD, J.M., WILLIAMS, J.C., ACKER, W.K., JACOBI, C., JANJIC, N. and MCHENRY, C.S. (2002): DNA polymerase III holoenzyme from

- Thermus thermophilus identification, expression, purification of components, and use to reconstitute a processive replicase. *Journal of Biological Chemistry* **277**(16): 13401-8.
- BURGERS, P.M., KOONIN, E.V., BRUFORD, E., BLANCO, L., BURTIS, K.C., CHRISTMAN, M.F., COPELAND, W.C., FRIEDBERG, E.C., HANAOKA, F., HINKLE, D.C., LAWRENCE, C.W., NAKANISHI, M., OHMORI, H., PRAKASH, L., PRAKASH, S., REYNAUD, C.A., SUGINO, A., TODO, T., WANG, Z., WEILL, J.C. and WOODGATE, R. (2001): Eukaryotic DNA polymerases: proposal for a revised nomenclature. *Journal of Biological Chemistry* **276**(47): 43487-90.
- DALRYMPLE, B.P., KONGSUWAN, K., WIJFFELS, G., DIXON, N.E. and JENNINGS, P.A. (2001): A universal protein-protein interaction motif in the eubacterial DNA replication and repair systems. *Proceedings of the National Academy of Sciences U S A* **98**(20): 11627-32.
- DERVYN, E., SUSKI, C., DANIEL, R., BRUAND, C., CHAPUIS, J., ERRINGTON, J., JANNIERE, L. and EHRLICH, S.D. (2001): Two essential DNA polymerases at the bacterial replication fork. *Science* **294**(5547): 1716-9.
- HOSFIELD, D.J., MOL, C.D., SHEN, B. and TAINER, J.A. (1998): Structure of the DNA repair and replication endonuclease and exonuclease FEN-1: coupling DNA and PCNA binding to FEN-1 activity. *Cell* **95**(1):135-46.
- HUGHES, A.J. JR, BRYAN, S.K., CHEN, H., MOSES, R.E. and MCHENRY, C.S. (1991): *Escherichia coli* DNA polymerase II is stimulated by DNA polymerase III holoenzyme auxiliary subunits. *Journal of Biological Chemistry* **266**(7):4568-73.
- INOUE, R., KAITO, C., TANABE, M., KAMURA, K., AKIMITSU, N. and SEKIMIZU, K. (2001): Genetic identification of two distinct DNA polymerases, DnaE and PolC, that are essential for chromosomal DNA replication in *Staphylococcus aureus*. *Mol Genet Genomics* **266**(4): 564-71.
- JERUZALMI, D., YURIEVA, O., ZHAO, Y., YOUNG, M., STEWART, J., HINGORANI, M., O'DONNELL, M. and KURIYAN, J. (2001): Mechanism of processivity clamp opening by the delta subunit wrench of the clamp loader complex of *E. coli* DNA polymerase III. *Cell* **106**(4): 417-28.
- KATO, J. and KATAYAMA, T. (2001): Hda, a novel DnaA-related protein, regulates the replication cycle in *Escherichia coli*. *EMBO Journal* **20**(15): 4253-62.
- KLEMPERER, N., ZHANG, D., SKANGALIS, M. and O'DONNELL, M. (2000): Cross-utilization of the beta sliding clamp by replicative polymerases of evolutionary divergent organisms. *Journal of Biological Chemistry* **275**(34): 26136-43.
- KONG, X.P., ONRUST, R., O'DONNELL, M. and KURIYAN, J. (1992): Three-dimensional structure of the beta subunit of *E. coli* DNA polymerase III holoenzyme: a sliding DNA clamp. *Cell* **69**(3): 425-37.
- KOONIN, E.V. and BORK, P. (1996): Ancient duplication of DNA polymerase inferred from analysis of complete bacterial genomes. *Trends in Biochemical Sciences* **21**(4): 128-9.
- LENNE-SAMUEL, N., WAGNER, J., ETIENNE, H. and FUCHS, R.P. (2002): The processivity factor beta controls DNA polymerase IV traffic during spontaneous mutagenesis and translesion synthesis in vivo. *EMBO Reports* **3**(1): 45-9.
- LOPEZ DE SARO, F.J. and O'DONNELL, M. (2001): Interaction of the beta sliding clamp with MutS, ligase, and DNA polymerase I. *Proceedings of the National Academy of Sciences U S A* **98**(15): 8376-80.
- LOW, R.L., RASHBAUM, S.A. and COZZARELLI, N.R. (1976): Purification and characterization of DNA polymerase III from *Bacillus subtilis*. *Journal of Biological Chemistry* **251**(5): 1311-25.
- NOIROT-GROS, M.F., DERVYN, E., WU, L.J., MERVELET, P., ERRINGTON, J., EHRLICH, S.D. and NOIROT, P. (2002): An expanded view of bacterial DNA replication. *Proceedings of the National Academy of Sciences U S A* **99**(12): 8342-7.
- SHAMOO, Y. and STEITZ, T.A. (1999): Building a replisome from interacting pieces: sliding clamp complexed to a peptide from DNA polymerase and a polymerase editing complex. *Cell* **99**(2): 155-66.
- SUTTON, M.D., NARUMI, I. and WALKER, G.C. (2002): Posttranslational modification of the umuD-encoded subunit of *Escherichia coli* DNA polymerase V regulates its interactions with the beta processivity clamp. *Proceedings of the National Academy of Sciences U S A* **99**(8): 5307-12.
- SUTTON, M.D., OPPERMAN, T. and WALKER, G.C. (1999): The *Escherichia coli* SOS mutagenesis proteins UmuD and UmuD' interact physically with the replicative DNA polymerase. *Proceedings of the National Academy of Sciences U S A* **96**(22): 12373-8.
- TANG, M., PHAM, P., SHEN, X., TAYLOR, J.S., O'DONNELL, M., WOODGATE, R. and GOODMAN, M.F. (2000): Roles of *E. coli* DNA polymerases IV and V in lesion-targeted and untargeted SOS mutagenesis. *Nature* **404**(6781): 1014-8.
- WAGNER, J., FUJII, S., GRUZ, P., NOHMI, T. and FUCHS, R.P. (2000): The beta clamp targets DNA polymerase IV to DNA and strongly increases its processivity. *EMBO Reports* **1**(6): 484-8.
- WARBRICK, E. (2000): The puzzle of PCNA's many partners. *BioEssays* **22**(11): 997-1006.
- YANG, W. (2000): Structure and function of mismatch repair proteins. *Mutation Research* **460**(3-4): 245-56.