Nucleic Acids Research, 1999, Vol. 27, No. 5 **1223–1242**

SURVEY AND SUMMARY

Conserved domains in DNA repair proteins and evolution of repair systems

L. Aravind[1,2], D. Roland Walker[1,3] and Eugene V. Koonin[1,*]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA, [2]Department of Biology, Texas A&M University, College Station, TX 70843, USA and [3]Department of Biology, Johns Hopkins University, Baltimore, MD 21218, USA

Received August 16, 1998; Revised and Accepted October 26, 1998

ABSTRACT

A detailed analysis of protein domains involved in DNA repair was performed by comparing the sequences of the repair proteins from two well-studied model organisms, the bacterium *Escherichia coli* and yeast *Saccharomyces cerevisiae*, to the entire sets of protein sequences encoded in completely sequenced genomes of bacteria, archaea and eukaryotes. Previously uncharacterized conserved domains involved in repair were identified, namely four families of nucleases and a family of eukaryotic repair proteins related to the proliferating cell nuclear antigen. In addition, a number of previously undetected occurrences of known conserved domains were detected; for example, a modified helix–hairpin–helix nucleic acid-binding domain in archaeal and eukaryotic RecA homologs. There is a limited repertoire of conserved domains, primarily ATPases and nucleases, nucleic acid-binding domains and adaptor (protein–protein interaction) domains that comprise the repair machinery in all cells, but very few of the repair proteins are represented by orthologs with conserved domain architecture across the three superkingdoms of life. Both the external environment of an organism and the internal environment of the cell, such as the chromatin superstructure in eukaryotes, seem to have a profound effect on the layout of the repair systems. Another factor that apparently has made a major contribution to the composition of the repair machinery is horizontal gene transfer, particularly the invasion of eukaryotic genomes by organellar genes, but also a number of likely transfer events between bacteria and archaea. Several additional general trends in the evolution of repair proteins were noticed; in particular, multiple, independent fusions of helicase and nuclease domains, and independent inactivation of enzymatic domains that apparently retain adaptor or regulatory functions.

INTRODUCTION

The DNA-based information system of most biological replicators present in the extant world is plagued by the possibility of insult from mutation. Given the vast number of mutagens present in the environment throughout the history of life, as well as the intrinsic error rate of DNA replication, one would imagine a strong selection for systems capable of safeguarding the genetic information. Indeed, the genomes of all cellular lifeforms and several large DNA viruses encode multiple proteins whose function is to repair the damaged DNA (1). In spite of the critical need for DNA repair, 'evolvability', that is, the ability to generate a certain level of uncorrected mutations, also seems to be selected for in the course of evolution. Organisms with an optimal level of evolvability have the best chance to survive environmental changes by virtue of stochastic variations in their genome, which provides the new raw material for natural selection. The complex interplay between the two opposing forces, namely the need for fidelity of transmission of genetic information and the need for evolvability, seem to define the organization of the repair systems.

DNA repair as a whole is a highly complex phenomenon. The repair mechanisms can be classified into several distinct, if not completely independent, major pathways that differ with regard to the level at which the lesions in damaged DNA are reversed or removed by the repair machinery: (i) direct damage reversal (DDR); (ii) base excision repair (BER); (iii) nucleotide excision repair (NER); (iv) mismatch repair (MMR); and (v) recombinational repair (RER). The general picture is further complicated by the existence of specialized, regulated forms of repair, such as the SOS response in bacteria, and by the intimate connection between repair, chromatin dynamics and the cell cycle in eukaryotes.

With the recent accumulation of complete genome sequences, it has become possible to systematically compare the repair systems of the respective organisms. Preliminary comparisons of this kind immediately made it clear that the repair machinery shows considerable variability, in terms of the present and absent genes, even in relatively close bacteria, such as *Escherichia coli* and *Haemophilus influenzae* (2). It was of major interest, therefore, to perform a systematic comparative analysis of the genes encoding

*To whom correspondence should be addressed. Tel: +1 301 435 5913; Fax: +1 301 480 9241; Email: koonin@ncbi.nlm.nih.gov

proteins involved in repair in the three superkingdoms of life—bacteria, archaea and eukaryotes—and in the main bacterial subdivisions. Here we present the results of such an analysis and discuss several previously undetected conserved domains that were uncovered in the process, as well as functional and evolutionary implications of the phyletic distribution of various repair genes.

DNA repair systems and mechanisms have been described in a comprehensive monograph by Friedberg and co-workers (1) as well as in several more recent, excellent reviews dedicated to specific aspects of repair (3–10). In this article, we make no attempt to cover the functional aspects of repair in any depth. Instead, we concentrate on those new facets of our understanding of the relationships between repair proteins and the evolution of repair systems that have been brought about by the comparative analysis of repair systems encoded in completely sequenced genomes. Whenever available, review articles are cited, and experimental work is cited only in as much as it is has a direct bearing on the conclusions drawn from genome analysis. Even with this focused approach, however, the number of relevant publications is quite substantial, and choices had to be made. We apologize to those researchers whose important work is not cited because of this, or simply by inadvertent but certainly regrettable omission.

APPROACH AND METHODS

Proteins were considered to be involved in DNA repair if on the basis of literature searches, they were found to meet one or more of the following criteria: (i) a role in repair demonstrated by genetic studies on model organisms, such as *E.coli* and the yeast *Saccharomyces cerevisiae*; (ii) a demonstrated role in human repair deficiency syndromes, such as Xeroderma pigmentosum, Cockayne syndrome, Bloom's syndrome, Werner's syndrome and allied diseases; (iii) possession of a biochemical activity compatible with a role in repair and the genetic data. The sequences of repair proteins from *E.coli* and yeast were subjected to detailed analysis with the SEALS package (11) which allows automated large-scale database searches using the PSI-BLAST program (12) after masking compositionally biased regions in the query sequences with the SEG program (13). The PSI-BLAST program uses the sequences retrieved from the database with a certain cut-off similarity level to construct a position-dependent weight matrix that is used for further iterations of the search, resulting in a significantly increased sensitivity and allowing the detection of subtle sequence similarities. During this iterative search, the random expectation (e) value computed by PSI-BLAST at the first instance when the given sequence is retrieved from the database is a reliable indication of the significance of a match, provided the low complexity regions in the query are appropriately masked. By default, each repair protein sequence from *E.coli* and yeast was compared to the non-redundant (NR) database at the National Center for Biteochnology Information (NIH, Bethesda) using PSI-BLAST run for three iterations. Further, case-by-case dissection of the protein families was performed where needed using PSI-BLAST searches run to convergence with the sequences of individual domains as queries as well as motif searches using the MoST program (14). Multiple alignments for the protein families were constructed using the –m4 option of PSI-BLAST, the CLUSTALW program (15) or the Gibbs sampling option of the MACAW program (16,17). Protein secondary structure predictions and structural database threading was performed using the PHD program (18,19). Structural models were manipulated using the Swiss-PDB –viewer program. The phyletic distribution of homologous proteins detected by the PSI-BLAST searches was assessed using the Tax_collector program of the SEALS package.

Throughout this analysis, an attempt was made to identify orthologous genes in different genomes. By definition, orthologs are genes (proteins) related by vertical descent or, in other words, direct evolutionary counterparts in different species. By contrast, paralogs have been defined as homologous genes derived by duplication within a species (20,21). This dichotomy does not fully describe the relationships between genes in distantly related genomes. Firstly, due to multiple lineage-specific gene duplications occurring subsequent to the radiation of the respective lineages, orthology generally cannot be described as a one-to-one relationship between these individual genes (22). Secondly, it is common in comparisons of proteins from phylogenetically distant species that the given domain architecture found in one of them has no counterpart in the other genome; instead, certain proteins from the second genome share a homologous domain(s) with the protein in question but otherwise have different domain organizations. Approaches for the identification of likely orthologs in genome comparisons have been described previously (22,23). Briefly, proteins or protein families from different genomes were considered orthologous if they showed the greatest similarity to each other among all proteins encoded by the two genomes and a similar (but not necessarily identical) domain architecture. We tried to distinguish, as clearly as possible, between apparent orthologs with similar domain organizations and non-orthologous proteins sharing one or more conserved domains. This distinction appears critical for reliable prediction of protein functions and for the construction of realistic evolutionary scenarios.

CONSERVED DOMAINS AND DOMAIN ARCHITECTURE IN DNA REPAIR PROTEINS

Escherichia coli and the yeast *S.cerevisiae* are the two model organisms in which DNA repair has been studied in most detail. The identified repair genes from these species were used as the basis for the comparative analysis of the domain architecture of repair proteins and the phyletic distribution of repair systems (Tables 1 and 2). The proteins comprising repair systems, like many other systems in the cell, appear to be designed according to a 'domain Lego' principle, that is by shuffling and recombining a limited repertoire of conserved domains (24–26). The nature of the domains is dictated by the activities required for repair, namely DNA binding, DNA strand cleavage, degradation and ligation, ATP-dependent duplex unwinding, and nucleotide polymerization. Accordingly, the main players in the repair systems are: (i) endo- and exonucleases and glycosidases, (ii) DNA helicases, (iii) ATPases (other than helicases) that are involved in such events as strand migration and loading of multiprotein repair complexes onto DNA, (iv) DNA ligases, (v) DNA polymerases and nucleotidyltransferases, (vi) DNA-binding domains and (vii) adaptors: protein–protein interaction domains that glue together diverse proteins in repair complexes and provide linkage to other cellular components, e.g. eukaryotic chromatin. Combined, nucleases and ATPases comprise the absolute majority of known DNA repair proteins (Tables 1 and 2).

Figure 1 shows the domain architectures of selected groups of DNA repair proteins. It appears that the combinations of helicases and polymerases with nuclease domains that have obvious utility in repair have been repeatedly invented in evolution as well as combination of each of these enzymes with distinct DNA-binding domains. By contrast, a helicase–polymerase combination is not common, but interestingly, it has been detected in a eukaryotic protein that is involved in DNA cross-link repair and whose domain architecture is conserved in eukaryotes (Fig. 1A; 27).

A major outcome of comparative sequence analysis is the delineation of novel conserved domains and prediction of their functions as well as discovery of new structural and evolutionary connections between previously identified domains. The sequences and subsequently structures of the main catalytic domains of polymerases, helicases and other ATPases have been characterized in detail in previous studies, and are readily recognizable due to the conservation of diagnostic motifs (e.g. 28–30). Thus the current analysis did not significantly expand these protein superfamilies. An interesting finding, however, is that several well-characterized DNA repair proteins contain domains with statistically significant similarity to helicases but with disrupted functional motifs, which suggests that while retaining the overall structure typical of helicases, they do not possess enzymatic activity. Examples of such apparent inactivation of helicases in repair systems include bacterial RecC and AddB proteins, transcription-repair coupling factor (Mfd or TRCF) and eukaryotic ERCC4 (Fig. 1A). Similar disruption of ATPase motifs probably leading to inactivation was observed in the ATPases of the RecA superfamily and, as reported previously, in the case of the central domain of UvrA (31) of the ABC superfamily (Fig. 1B).

Nucleases generally tend to be less conserved in evolution than ATPases or polymerases. Some superfamilies, e.g. the 3′→5′ nucleases (32), the 5′→3′/FLAP nuclease superfamily (33), as well as the phosphoesterase superfamily that includes such nucleases as SbcD and Mre11 (34), have been extensively studied. There are, however, many other groups of nucleases that have not been characterized in comparable detail, and in the course of the present analysis, we have delineated four superfamilies of nucleases that to our knowledge, have not been recognized previously, and identified the likely origin of another major superfamily.

AP endonuclease/ENDO4 superfamily

Bacterial endonuclease IV is a homolog of eukaryotic apurinic endonucleases (35). Representatives of this family of endonucleases were detected in all bacterial, archaeal and eukaryotic species. Unexpectedly, iterative database searches revealed statistically significant similarity ($e \sim 10^{-4}$, iteration 3) between this endonuclease family and sugar isomerases (including xylose isomerases, tagatose epimerases and hexulose isomerases) that have the TIM barrel structural fold. The endonucleases and sugar isomerases share several conserved motifs, in particular the [DE]X2H signature as well as four histidines that are conserved in most of the proteins (Fig. 2A). Secondary structure-based threading and modeling of the AP endonuclease using the xylose isomerase structure (36,37) as the template indicate that they have similar structures, with the conserved histidines distributed in the interior of the TIM barrel (Fig. 3) and probably involved in metal coordination similarly to the deaminase-urease superfamily of TIM barrels (38). On the basis of this structural model, it can be predicted that in the AP endonucleases the deoxyribose of DNA is positioned in the active site similarly to the placement of xylose in the xylose isomerases (Fig. 3). Interestingly, the recently characterized new group of nucleases involved specifically in the repair of UV-damaged DNA [mus18/UVDE from *Neurospora* (39) and *Schizosaccharomyces* and their *Bacillus* ortholog YwjD (40)] was also found to belong to this superfamily of TIM barrel enzymes.

UvrC endonuclease superfamily (Uri domain)

UvrC protein is the endonuclease subunit of the bacterial excision repair complex that consists of the ABC-type ATPase UvrA and the helicase UvrB (41,42). Iterative database searches showed that UvrC contained a domain with statistically significant similarity ($e < 10^{-3}$ at the sixth iteration) to intron-encoded endonucleases and several uncharacterized bacterial, archaeal and viral proteins (we designated this domain Uri after UvRC and Intron-encoded endonucleases). This previously undetected endonuclease family contains a RX_3[YH] sequence signature, two conserved tyrosines that typically are separated by 10 residues, and a conserved glutamate (Fig. 2B). These conserved polar residues likely participate in catalysis and, indeed, the role of the conserved arginine in the activity of the intron-encoded endonucleases has been demonstrated by site-directed mutagenesis (43). A highly conserved group of small, functionally uncharacterized proteins from different bacteria, eukaryotes and viruses belong to this superfamily of nucleases and may have as yet unknown roles in repair. Another subfamily of putative nucleases that belongs to this family is highly conserved in archaea and contains a C-terminal metal-binding cluster that may be involved in DNA binding. Interestingly, in an uncharacterized mycobacterial protein, the Uri nuclease domain is fused to a 3′–5′ exonuclease domain homologous to the ε subunits of PolIII (e.g. *E.coli* DnaQ), whereas in the archaeon *Methanococcus jannaschii*, a UvrC-endonuclease III fusion was detected (Fig. 2B).

EndoV endonuclease superfamily

The endonuclease V (*E.coli nfi* gene product), which is highly conserved in eukaryotes, showed subtle but statistically significant similarity ($e < 10^{-3}$ in the second PSI-BLAST iteration) to a region of UvrC that is located between the Uri domain and the C-terminal helix–hairpin–helix (HhH) domain. Multiple alignment of the EndoV family with the UvrC sequences showed the conservation of two aspartates and a lysine that may be directly involved in catalysis as well as several potential structural elements (Fig. 2C). The site-directed mutagenesis results on UvrC (42) not only confirm the essential role of the two conserved aspartates but also help delineate the exact role of the two nuclease domains of UvrC in NER repair. UvrABC removes a patch of DNA around a lesion by making two incisions at both sides of a modified base, namely 8 nt 5′ and 15 nt 3′ (41,42). Mutation of the conserved D399 and D466 in *E.coli* UvrC (Fig. 2C) abolished the 5′ incision but did not effect the 3′ incision (42). Thus it can be confidently predicted that the EndoV domain catalyzes the 5′ incision, whereas the Uri domain is responsible for the 3′ incision.

Table 1. *Escherichia coli* DNA repair systems: conservation in completely sequenced genomes

Protein	Function / Activity	Pathway[a]	Phylogenetic distribution[b]						Domains[d]	Comment
			Bacteria[c]				Archaea	Eukarya		
			PB	G+	CB	SP				
PhrB	Photolyase	DR	+	+	+	-	+	+	Flavin and 8-hydroxy-5-deazaflavin- dependent light receptor domain	Markedly episodic distribution; among G+, found in *B. firmus* and *Streptomyces* but not in B. subtilis or *Mycobacteria*; three copies in *Synechocystis* (see text).
Ada	O-6 alkylguanine, O-4 alkylthymine alkyltransferase; removes alkyl groups of many types; transcription activator	DR	+	+	-	-	(+)	(+)	C2C2 Zn finger+AraC family HTH+ methyltransferase	Archaeal and eukaryotic homologs have only the methyltransferase domain; paralog of Ogt
Ogt	O-6-methylguanine DNA methyltransferase	DR	+	+	-	-	+	+	methyltransferase	Paralog of Ada without the additional domains
MutT	8-oxo-dGTPase	DR	+	+	+	-	(+)	(+)	MutT (Nudix) hydrolase	A vast family of pyrophosphohydrolases; some of the orthologous relationships should be considered provisional
Dut	dUTPase	DR	+	+	-	+	-	+	dUTPase domain distantly related to dCTP deaminase (*E. coli* Dcd) (151)	Among the spirochaetes, found in *T. pallidum* but not *B. burgdorferi;* also encoded by several bacteriophages, poxviruses, and herpesviruses
Dcd	dCTP deaminase	DR	+	+	+	-	+	-	Dcd/Dut domain	Paralog of Dut. Universal in Archaea but episodic in bacteria
AlkA	3-methyladenine, 3-methylguanine, O-2-methylcytosine, O-2-methyl thymine DNA glycosylase II	DR, BER	+	+	+	(+)	+	+	Glycosidase+HhH	Family of α helical glycosidases homologous to endonuclease III (Nth, MutY). The Mycobacterial ortholog has an N-terminal fusion of an Ada-like C2C2 Zn finger
AlkB	Unknown	DR, BER(?)	+	-	-	-	-	+	Novel predicted hydrolase domain	New family found only in E. coli and Caulobacter, and in a diverged form in Mycobacterium among bacteria, but also in animals and plants (but not in yeast) and in the polyproteins of plant RNA viruses of the carla- and trichoviruses groups (L. Aravind and E. V. Koonin, unpublished observations)
MutY	8-oxoguanine DNA glycosylase & AP-lyase, A-G mismatch DNA glycosylase	BER, MMY	+	+	+	+	+	+	Glycosidase/endonuclease+HhH+ cysteine-rich motif also seen in the C-terminus of some RecB family nucleases	Family of α helical glycosidases/endonucleases (MutY, nth, and AlkA are paralogs)
Nth	Endonuclease III & thymine glycol DNA glycosylase	BER	+	+	+	+	+	+	Glycosidase/endonuclease+HhH+ cysteine-rich motif also seen in at the C-terminus of some RecB family nucleases	Family of α helical glycosidases/endonucleases (MutY, nth, and AlkA are paralogs)
MutM/ Fpg	Formamidopyrimidine & 8-oxoguanine DNA glycosylase	BER	+	+	+	-	-	-	Distinct Glycosidase/endonuclease+HhH+C4 little finger motif	Paralog of Nei
Nei	Endonuclease VIII	BER	+	+	+	-	-	-	Distinct Glycosidase/endonuclease+HhH+C4 little finger motif	Paralog of MutM
Nfo	Endonuclease IV	BER	+	+	-	-	+	+	Endonuclease of the ENDOIV/AP superfamily	See text and Fig. 2A
Nfi (YjaF)	Endonuclease V	BER	+	+	-	-	+	+	UvrC-like 3'-incision endonuclease domain	See text and Fig. 2C
PolA	DNA polymerase I	BER	+	+	+	+	-	+	RNAaseH-type 5'-3' exonuclease+3'-5' exonuclease + polymerase	The eukaryotic orthologs have a N- terminal SFII helicase fusion (see text and Fig. 1A)
Tag	3-methyladenine DNA glycosylase I	BER	+	+	-	-	-	-	Distinct glycosidase domain	So far found only in Proteobacteria and Mycobacteria
Ung	uracil DNA glycosylase	BER	+	+	-	+	-	+	Distinct glycosylase domain	Also found in herpesviruses and poxviruses
XthA	Exodeoxyribonuclease III	BER	+	+	+	+	+	+	Sphingomyelinase-DNAase	Homolog of LINE retroposon endonucleases; universal except *Mycoplasma*.
RadC	Predicted DNA-binding protein	BER	+	+	+	-	-	-	HhH+uncharacterized conserved domain	*E. coli* also encodes 3 paralogs of RadC (YkfG, YfjY and YeeS that lack the HhH domain)
RadA / Sms	Predicted ATP-dependent protease	NER, BER	+	+	+	-	-	-	C4 Zn finger-like domain +RecA family ATPase+ Lon-like protease	The protease appears to be active in some forms of the protein and inactive in others. Stand-alone forms of the protease domain are found in other proteins.
Mfd	transcription repair coupling factor; helicase	NER	+	+	+	+	-	-	SFII helicase(disrupted) + SFII helicase	Fusion with a disrupted uvrB-like helicase in the N- terminal. Universal in bacteria except Mycoplasma
UvrA	ATPase, DNA binding	NER	+	+	+	+	+	-	ABC family ATPase+Finger (see Fig. 1B)	Universal in bacterai; among the archaea, found only in *Methanobacterium*
UvrB	Helicase	NER	+	+	+	+	+	-	SFII helicase+ URBC domain	Possibly intereacts with UvrC with the common URBC domain.
UvrC	Nuclease	NER	+	+	+	+	+	-	5'-endonuclease+3'-endonuclease+HhH+UVRBC	See text and Fig. 2B,C
UvrD	helicase II; initiates unwinding from a nick	NER, mMM, SOS	+	+	+	+	-	-	SFI helicase	Universal in bacteria
MutL	predicted ATPase	mMM, VSP	+	+	+	+	-	+	HSP90 family ATPase	ATPase of the HSP90-gyrase family
MutS	ATPase	mMM, VSP	+	+	+	+	+	+	ABC superfamily ATPase	2 distinct subfamilies in bacteria and several in the eukaryotes. Among the archaea, only in *Methanobacterium* and *Pyrococcus*.
MutH	Endonuclease	mMM	+	-	-	-	-	-	Sau3-like restriction endonuclease domain	So far detected only in *E. coli*
Dam	GATC-specific N-6 adenine methlytransferase; imparts strand specificity to mismatch repair.	mMM	+	+	+	-	+	-	Adenine-specific DNA methylase	Among Archaea, only in *M. jannaschii*
Vsr	strand-specific, site specific, GT mismatch endonuclease; fixes deamination resulting from Dcm	VSP	+	-	+	-	-	-		Nostoc is the only known Cyanobacterium with this domain so far.
XseA / nec7	Exonuclease VII, large subunit	MM	+	+	-	-	-	-	A distinct nuclease domain	No detectable relationship with other nucleases and very limited distribution
XseB	Exonuclease VII, small subunit	MM	+	+	-	-	-	-	Uncharacterized domain	
SbcB	Exodeoxyribonuclease I	mMM, RER	+	-	-	-	-	-	3'-5' Exonuclease fold	A highly divergent version of the domain so far detected only in *E. coli* and *H. influenzae*
Dcm	site-specific C-5 cytosine methlytransferase; VSP is targeted toward hotspots created by dcm	mMM	+	+	+	-	+	-	SAM-dependent methyltransferase	
DinP	Specific function unknown (predicted nucleotidyltranferase)	MM, RER	+	+	-	-	+	+	Nucleotidyl transferase+HhH	Paralog of UmuC. Among archaea, so far only in *Sulfolobus*
SbcC	exonuclease subunit, predicted ATPase	RER	+	+	+	+	+	+	ABC family ATPase with coiled coils	Nearly universal but missing in *Mycoplasmas, H. influenzae, H. pylori;* in spite of the preponderance of the coiled-coil structure, orthology could be shown through distinct signature motifs
SbcD	Exonuclease	RER	+	+	+	+	+	+	Calcineurin-like phosphohydrolase domain	In spite of the large number of superfamily members, orthologous relationships between repair enzymes are apparent; missing in *Mycoplasmas, H. influenzae, H. pylori.*

Table 1. *Continued*

RecA	recombinase; ssDNA-dependent ATPase, activator of LexA autoproteolysis	RER, SOS	+	+	+	+	+	+	RecA/Sms family ATPase	Eukaryotic and archaeal proteins contain in addition an N-terminal HhH-like domain (see text and Fig. 1B)
RecB	Helicase/exonuclease	RER	+	+	(+)	+	(+)	(+)	SFI helicase + nuclease domain (see text and Fig. 1A, 2D)	Among G+, only *Mycobacterium* has a true ortholog; *B. subtilis* has a paralog (AddA). Archaea and eukaryotes have only distantly related helicases and nucleases.
RecC	Helicase/exonuclease	RER	+	+	(+)	+	-	-	Disrupted SFI helicase with intact C- terminal 2 motifs.	Among G+, only *Mycobacterium* has a true ortholog; *B. subtilis* has a distant paralog (AddB)
RecD	Helicase/exonuclease	RER	+	+	(+)	+	(+)	-	Helicase.	Needed for maximal activity of the recBCD nuclease, though not an active helicase. Among G+, only *Mycobacterium* has a true ortholog; among archaea, only *M. jannaschii* has a highly similar homolog
RecF	predicted ATPase; required for resumption of DNA replication at disrupted replication forks	RER	+	+	+	+	(+)	(+)	ABC superfamily ATPase with coiled coil regions.	
RecG	Holliday junction-specific DNA helicase; branch migration inducer	RER	+	+	+	+	(+)	(+)	HhH + SFII Helicase	Orthologs in all bacteria except *Mycoplasma*
RecJ	Nuclease	RER	+	+	+	+	+	-	DHH domain nuclease	Orthologs in all bacteria except *Mycoplasma* (95)
RecN	predicted ATPase	RER	+	+	(+)	+	(+)	(+)	ABC superfamily ATPase with coiled coil regions.	
RecO	"anti-ssb factor"; stabilization of RecA filaments; ATP-independent, RecA-like strand assimilation activity	RER	+	+	+	-	-	-	Conserved N-terminal domain but variable C- terminal domains; a C-terminal Zn finger in *B. subtitlis*	
RecQ	helicase; suppressor of illegitimate recombination	RER	+	+	+	+	(+)	+	Helicase+HRD domain	See text and Fig. 1A
RecR	required for resumption of DNA replication at disrupted replication forks	RER	+	+	+	+	-	-	HhH+C4 finger+Toprim(inactive)	2 paralogs in G+ (RecR and RecM)
RusA (YbcP)	endonuclease /Holliday junction resolvase	RER	+	+	-	-	-	-	HhH motif	Detected only in *E. coli*, *B. subtilis*, *Aquifex*, and bacteriophages; horizontal transfer likely
RuvA	Holliday junction resolvase	RER	+	+	+	+	-	-	HhH is the only detectable motif	Universal in bacteria
RuvB	Holliday junction resolvase; ATPase subunit of a helicase,	RER	+	+	+	+	-	-	AAA family ATPase	Universal in bacteria
RuvC	Holliday junction resolvase; endonuclease	RER	+	+	+	+	-	-	RNAaseH-llike nuclease	
RecE	exonuclease VII	RER	+	-	-	-	-	-	C-terminal RecB-like nuclease domain.	
RecT	annealing protein	RER	+	+	-	-	-	-	Unique domain	So far detected only in *E. coli*, *B. subtilis*, and phage SPP1
DinG	predicted helicase; SOS inducer	SOS	+	+	(+)	(+)	(+)	(+)	SFII, Snf/Swi Helicase, in G+ fused with 3'-5' exonuclease	Archaea and eukaryotes have only distantly related Snf/Swi helicases
LexA	transcriptional regulator, autoprotease	SOS	+	+	+	-	-	-	HTH+ signal peptidase type β- meander domain	The protease domain is related to the signal peptidases and to other non-protease β meander proteins
PolB	DNA polymerase II	SOS	+	-	-	-	+	+	Family B polymerase	Among bacteria, found only in *E coli;* possible *gene* transfer from a bacteriophage.
UmuC	in conjunction with umuD and recA, facilitates translesion DNA synthesis	SOS	+	+	+	-	(+)	(+)	Nucleotidyl transferase+HhH	Among archaea, so far found only in *Sulfolobus*; Some of the eukaryotic paralogs also contain BRCT and HhH domains (Fig. 1C).
UmuD	in conjunction with umuC and recA, facilitates translesion DNA synthesis; autoprotease	SOS	+	(+)	(+)	(+)	-	-	HTH+ signal peptidase type β- meander domain	A paralog of LexA so far found only in Enterobacteria
DnaE	polymerase subunit of the DNA polymerase III holoenzyme	MP	+	+	+	+	-	-	PHP(predicted phosphatase)+Polymerase catalytic domain+HhH	DNA polymerase III α subunit. Universal in bacterial, 2-3 members in G+ (33)
DnaQ	3'-5' exonuclease subunit of the DNA polymerase III holoenzyme	MP	+	+	+	+	+	-	3'-5' Exonuclease	Fused to the DNA polymerase III α subunit (DnaE) in G+. Among the Archaea, only in Archaeoglobus.
DnlJ	DNA ligase	MP	+	+	+	+	-	-	NAD dependent ligase+HhH+BRCT	Universal, with conserved domain architecture, in bacteria
Ssb	Single-strand binding protein	MP	+	+	+	+	-	+	OB-fold-like domain	Mitochondrial protein in eukaryotes

[a]DR, damage reversal; BER, base excision repair; NER, nucleotide excision repair; MM, mismatch repair; mMM, methylation-dependent mismatch repair; MMY, mutY-dependent mismatch repair; VSP, very short patch mismatch repair; RER, recombinational repair; MP, multiple pathways.

[b]Evolutionary relationships were defined with respect to the *E.coli* proteins.

+ indicates the presence of an apparent ortholog with partially, if not completely, conserved domain architecture, in at least one representative of the given lineage (additional details of the distribution are given in the Comments column); (+) indicates the presence of a non-orthologous homolog, typically with a significantly different domain architecture, but with at least one homologous domain with highly significant sequence similarity; – indicates the absence of homologs. In some cases, there was a degree of arbitrariness in assigning the '(+)' or the '–' status. For example, in the broadest sense, all helicases and ATPases are represented by homologs, even if very distant ones, in all lineages. However, in order to emphasize the distinction between specific and very general relationships, the (+) status was assigned only in cases when a representative of the particular family, to which a given *E.coli* ATPase belongs, was detected in the specific lineage.

[c]PB, proteobacteria; G+, Gram-positive bacteria; CB, cyanobacteria; SP, spirochaetes.

[d]The 'domains' here are defined operationally as conserved parts of proteins that have the potential to exchange and appear to evolve independently. They frequently but not necessarily correspond to actual structural domains. The domains are indicated according to their linear order in the protein sequence, from N- to C-terminus.

RAD1/ERCC4 endonuclease superfamily and its inactivated derivatives

The human ERCC4 protein and its yeast ortholog RAD1 are endonucleases involved in NER (44). Our analysis revealed orthologs of this enzyme in archaea but not in bacteria. Additionally, a second paralog of ERCC4 was detected in the genomes of *S.cerevisiae* and *Schizosaccharomyces pombe* and may belong to a novel eukaryotic repair pathway. The only detectable bacterial member of this family is an uncharacterized protein from *Mycobacterium tuberculosis*. All the (predicted) nucleases of this superfamily contain the strikingly conserved signature $ERKX_2SD$ as well as an additional conserved aspartate; the conserved negatively-charged residues are likely to function in metal ion coordination and as nucleophiles in catalysis (Fig. 2D). Most of the repair proteins containing this type of nuclease have a distinct domain organization, with an N-terminal superfamily 2 helicase domain, followed by the nuclease domain and the C-terminal DNA-binding HhH domains (45) (Fig. 1A). The remarkable feature of this protein family is that in archaea,

Table 2. Yeast DNA repair systems: conservation in completely sequenced genomes

Gene[a]	Function / Activity	Pathway	Organisms[a] Ce	Hs	Arc	Bac	Domains	Comments
RAD3 epistasis group								
RAD1	Single-strand DNA endonuclease; Cuts at duplex/3'single-strand junctions	NER, RER	+	+	+	(+)	Disrupted SFII helicase+ nuclease (ERCC4 family)+HhH	See text and Figs. 1A and 2E
RAD10	Single-strand DNA endonuclease subunit (RAD1-RAD10 complex)	NER, RER	+	+	-	-	Disrupted nuclease (ERCC4 family)+ HhH	The yeast protein lacks the C-terminal HhH domains found in the orthologs from other species including *S. pombe*
RAD2	Single-strand DNA endonuclease/5'-3' exonuclease; cleaves at duplex / 5' single-strand junctions	NER	+	+	+	+	5'-3 exonuclease+HhH	Some of the eukaryotic members contain large non-globular inserts. In bacteria, the orthologous domain is fused to DNA polymerase I.
RAD3	5'-3' DNA and DNA-RNA helicase; Pol II basal transcription factor	NER	+	+	(+)	(+)	SFII helicase	DEAH family, SFII helicase
RAD4	Possibly involved in repair-transcription coupling and in the repair-cell cycle connection; exact role unknown	NER	-	+	-	-	No known domains detectable	XP-C ortholog. Yeast also has a paralogous gene which is conserved in *S. pombe*
RAD7	Involved in NER as a complex with RAD16	NER	(+)	-	-	-	Leucine-rich repeats	Leucine rich repeat protein.
RAD16	DNA helicase	NER	+	+	(+)	(+)	SNF/SWI helicase+RING finger (insert between SF2 motifs 4 and 5)	All eukaryotic orthologs have a RING finger
RAD14	Damage-specific DNA-binding protein	NER	+	+	-	-	C2C2 Zn finger + H2C2 finger-like motif	
RAD23	Provides connection between NER and ubiquitin-dependent proteasome pathways	NER	+	+	-	-	Ubiquitin+uncharacterized conserved domain	
RAD25 (SSL2, UVS12)	3'-5' DNA helicase; Pol II basal transcription protein	NER	+	-	(+)	+	Uncharacterized conserved domain+SFII helicase	Superfamily II helicase with a specific conserved N- terminal domain
SSL1	TFIIH 44 kD subunit	NER	+	+	-	-	von Willebrand factor A domain + Zn finger.	Homolog of the proteasomal subunit S5(152, and L. Aravind, unpublished observations)
TFB1	TFIIH 62 kD subunit	NER	+	+	-	-	Novel repetitive motif	
CDC9	DNA Ligase	NER	+	+	(+)	(+)	ATP-dependent ligase + BRCT	Archaeal and bacterial ATP-dependent ligases lack the BRCT domain
MMS19	Transcription/repair protein apparently acting through interaction with TFIIH	NER	+	-	-	-	Leucine-rich repeats	
SNM1 (PSO2)	Protein required for DNA cross-link repair; predicted nuclease	NER	+	+	(+)	(+)	Metallo β-lactamase	
RAD6 epistasis group								
RAD5(REV2)	Helicase	TLR[b]	+	+	(+)	(+)	SNF/SWI helicase+RING finger	Domain architecture analogous to

the N-terminal domain contains intact conserved superfamily II helicase motifs and is predicted to be an active helicase, whereas in eukaryotes, this domain appears to be inactivated, as indicated by the disruption of the helicase motifs (Fig. 2D). The archaeon *Archaeoglobus fulgidus* and African swine fever virus encode smaller proteins that seem to consist only of the nuclease domain and the HhH domain (Fig. 1A).

Further iterative database searches using the nuclease-HhH portion of the ERCC4 family proteins as the query detected a relationship with another family of eukaryotic repair proteins that includes human ERCC1 and its homologs in other eukaryotes, such as yeast RAD10 (Fig. 1A). The sequences of these proteins are similar to that of RAD1 at a statistically significant level ($e < 10^{-3}$ in the third iteration) but contain substitutions of some of the predicted catalytic residues, in particular the ERKx2SD motif, indicating that their nuclease domain is probably inactive (data not shown). Notably, yeast RAD1 functions as a stable complex with RAD10 (46).

The RecB nuclease domain family

The C-terminal portion of the RecB (*E.coli*) and AddA (*Bacillus subtilis*) subunits is required for the nuclease activity of the recBCD and AddABC complexes, respectively (47,48). Sequence analysis performed using PSI-BLAST showed that this domain is present as a stand-alone version in several bacterial, archaeal, eukaryotic and phage proteins, and also is fused to other superfamily I helicases such as yeast DNA helicase 2 and its orthologs from other eukaryotes, in which it is located N-terminal to the helicase domain, in contrast to its location in RecB and AddA (Fig. 1A). This putative nuclease domain was also detected in the C-terminal part of RecE, another repair nuclease from *E.coli*. On the basis of these observations, we propose that this novel nuclease domain tends to function in conjunction with superfamily I helicases and has been fused to them independently, on more than one occasion. Multiple alignment of this nuclease family shows the presence of [GV]hhD and [DE]hK (h indicates a hydrophobic residue) signatures and a conserved tyrosine near the C-terminus (Fig. 2E). Given the strict conservation of this tyrosine, it may be involved in the formation of a covalent intermediate with the cleaved DNA strand as shown for several classes of enzymes that catalyze DNA cleavage, such as Flp recombinases, topoisomerases and enzymes involved in rolling circle replication (49–52).

DNA-BINDING DOMAINS

All components of the DNA repair machinery must be delivered to the sites of their action on DNA—some bind DNA directly, whereas others rely on protein–protein interactions. Many repair

Table 2. *Continued*

							(insert between SF2 motifs 4 and 5)	RAD16 and other eukaryotic SNF/
RAD6	Ubiquitin conjugating enzyme, connects repair with protein degradation; forms DNA-binding heterodimers with RAD18	RER, TLR, and novel pathways	+	+	-	-	Ubiquitin conjugating enzyme	
RAD9	DNA damage checkpoint component	RER	-	-	-	-	2 C-terminal BRCT domains	So far no orthologs detectable
RAD18	Forms DNA-binding heterodimers with RAD6	TLR	(+)	(+)	-	-	RING finger+additional, finger-like conserved motif	Genuine orthologs seen only in other fungi.
RADH(SRS2)	DNA helicase involved in the RAD1-dependent NER pathway	NER, RER	(+)	(+)	(+)	(+)	SF1 helicase.	
REV1	DNA polymerase ζ subunit, predicted template-independent nucleotidyl transferase	TLR	+	(+)	(+)	(+)	Nucleotidyl transferase+ HhH+BRCT	Prokaryotic homologs, such as DinP, lack the BRCT domain but have HhH motifs.
REV3	DNA polymerase ζ elongation subunit	TLR	(+)	+	(+)	-	Superfamily B DNA polymerase + unique Cys-rich domain	
REV7	DNA polymerase ζ subunit	TLR	(+)	(+)	-	-	Horma domain	
RAD30	Novel rad6/18 dependent pathway component	Novel pathway	+	(+)	(+)	(+)	Nucleotidyl transferase	A distant paralog of REV1
PSO4(PRP19)	Connection between mRNA processing and repair	RER	+	+	-	-	WD40 repeats	
RAD52 epistasis group	Proteins of this group are primarily involved in the recombinational repair of double-strand breaks							
RAD50	Chromatin modifying ATPase	RER	+	+	+	+	ABC superfamily ATPase+ large, inserted coiled coil domains	
RAD51	ATPase	RER	+	+	+	+	Modified HhH+RecA-type ATPase	See text and Fig. 1B
RAD52	Forms a complex involved in strand exchange with RAD55 and RAD57	RER	-	+	-	-	Uncharacterized conserved domain	Paralog of RAD59
RAD53(SPK1)	Protein Ser/Thr kinase	RER	+	-	-	-	FHA+Ser/Thr kinase+FHA	
RAD54	Helicase involved in strand exchange in conjunction with RAD51	RER	+	+	(+)	(+)	SNF/SWI type SF2 helicase	
RAD55	ATPase	RER	-	+	+	-	Modified HhH+RecA-type ATPase	See text and Fig. 1B
RAD57	ATPase	RER	+	+	+	+	Modified HhH+RecA-type ATPase	See text and Fig. 1B
RAD24	ATPase; DNA damage checkpoint component interacting with RAD17 and MEC3	RER	(+)	(+)	(+)	(+)	RF-C type AAA superfamily ATPase	
MRE11	3'-5' exonuclease and endonuclease; as a complex with RAD50, involved non-homologous joining of DNA ends	RER	+	+	+	+	Nuclease of the calcineurin-like fold	
RAD59	Involved in double-strand break repair, function unknown	RER	-	(+)	-	-	Uncharacterized conserved domain	Paralog of RAD52
XRS2	Involved in double-strand break repair as a complex with RAD50 and MRE11	RER	-	-	-	-	Divergent FHA domain + coiled coil.	
RAD17	DNAase; DNA damage checkpoint component interacting with RAD24 and MEC3	RER	(+)	+	(+)	(+)	PCNA fold domain	See text and Fig. 2F
MEC3	DNA damage checkpoint component interacting with RAD17 and RAD24		-	-	-	-		No identifiable structural features
MMS21	Function unknown	?	+	-	-	-	Zn-coordinating Cys-His cluster	
REC114	Function unknown	RER	-	-	-	-	Coiled-coil	No detectable homologs
REC102	Function unknown	RER	-	-	-	-	No identifiable structural features	No detectable homologs
REC103(Ski8)	Function unknown; predicted adaptor	RER	-	-	-	-	WD40 repeats	Ortholog in *S. pombe*
RER104	Function unknown	RER	-	-	-	-	No identifiable structural features	
RNC1	Claimed to be a nuclease (133)	RER	-	+	+	+	divergent S1 domain + SAM-dependent methyltransferase	
SPO11	Double-strand break introducing endonuclease	RER	+	-	+	-	Divergent Toprim domain (100)	The archaeal orthologs are subunits of Topoisomerase VI.
MEC1	DNA dependent protein kinase; DNA damage checkpoint component; phosphorylates DDC1	RER	+	+	-	-	Lipid kinase superfamily domain	
DDC-1	Checkpoint sensor	RER	-	-	-	-		Distantly related to Rad-9 from *S pombe*.
Other repair proteins								
OGG1	8-oxoguanine DNA glycosylase	DR, BER	-	+	+	+	glycosidase domain+HhH	Family of α helical glycosidases homologous to endonuclease III
NTG1	8-oxoguanine DNA glycosylase	DR, BER	+	+	+	+	glycosidase domain+HhH	Family of α helical glycosidases homologous to endonuclease III; close paralog of NTG2 and distant paralog of OGG1
NTG2	8-oxoguanine DNA glycosylase	DR, BER	+	+	+	+	glycosidase domain+HhH	Family of α helical glycosidases homologous to endonuclease III; close paralog of NTG1 and distant paralog of OGG1
PIF1	5'-3' helicase involved in mitochondrial repair	?	+	(+)	(+)	(+)	SF1 helicase	
RAD26	Helicase involved in transcription-repair coupling	NER	+	+	(+)	(+)	SNF/SWI type SF2 helicase	
KEM1(RAR5)	5'->3' Nuclease	RER	+	+	-	-	Novel nuclease domain	
RAD27	single-strand DNA endonuclease/5'-3' exonuclease	NER, MMR?	+	+	+	+	5'-3 exonuclease+HhH	
DIN-7	5'->3' Nuclease		+	+	+	+	5'-3 exonuclease+HhH	In bacteria, the orthologous domain is fused to DNA polymerase I.
EXO-1	5'->3' Nuclease	MMR	+	+	+	+	5'-3 exonuclease+HhH	In bacteria, the orthologous domain is fused to DNA polymerase I.
PHR1	Photolyase	DR	-	-	+	+	Flavin and 8-hydroxy-5-deazaflavin-dependent light receptor domain	

[a]Ce, *C.elegans*; Hs, *H.sapiens*; Arc, archaea; Bac, bacteria. See also footnotes to Table 1. [b]TLR, trans-lesion repair.

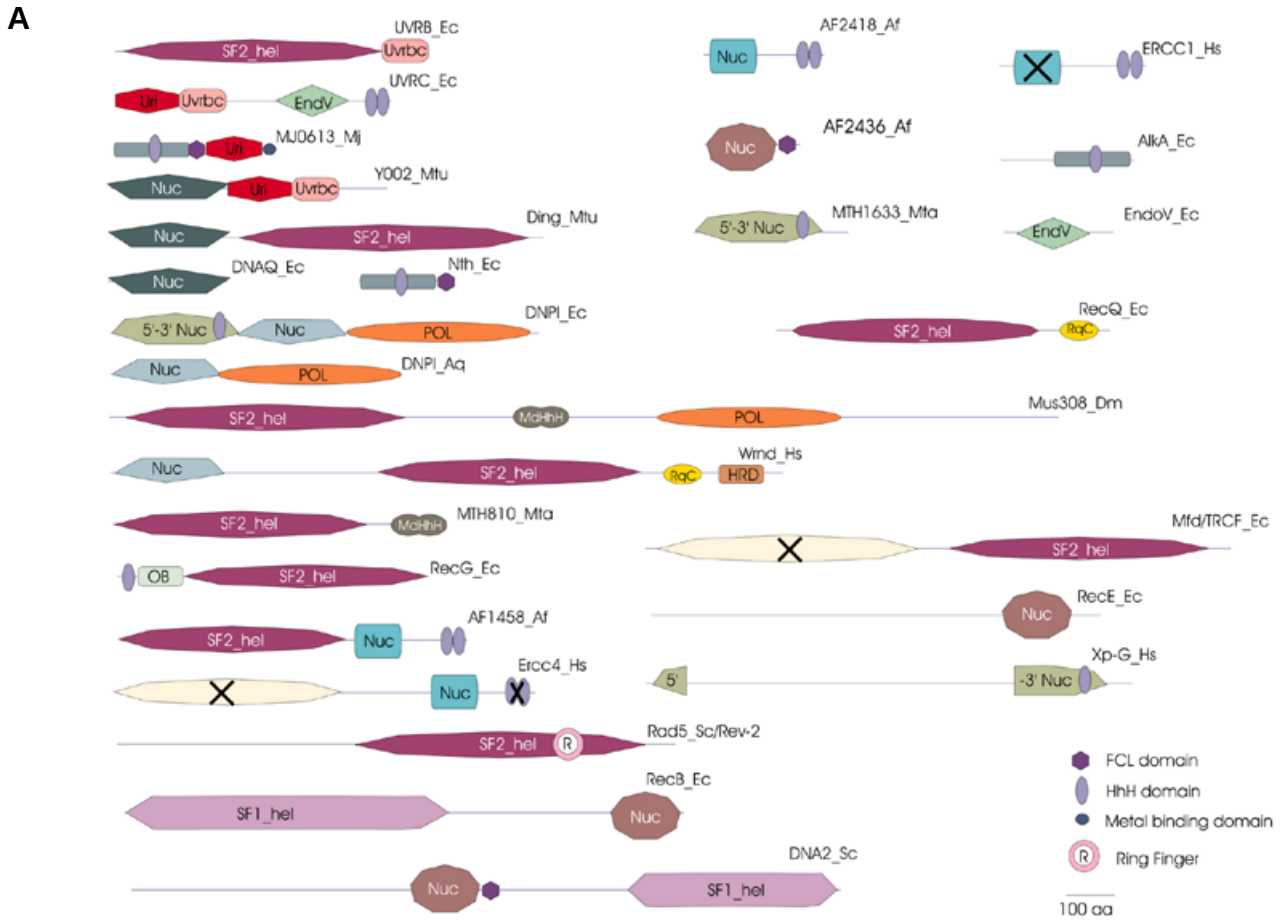
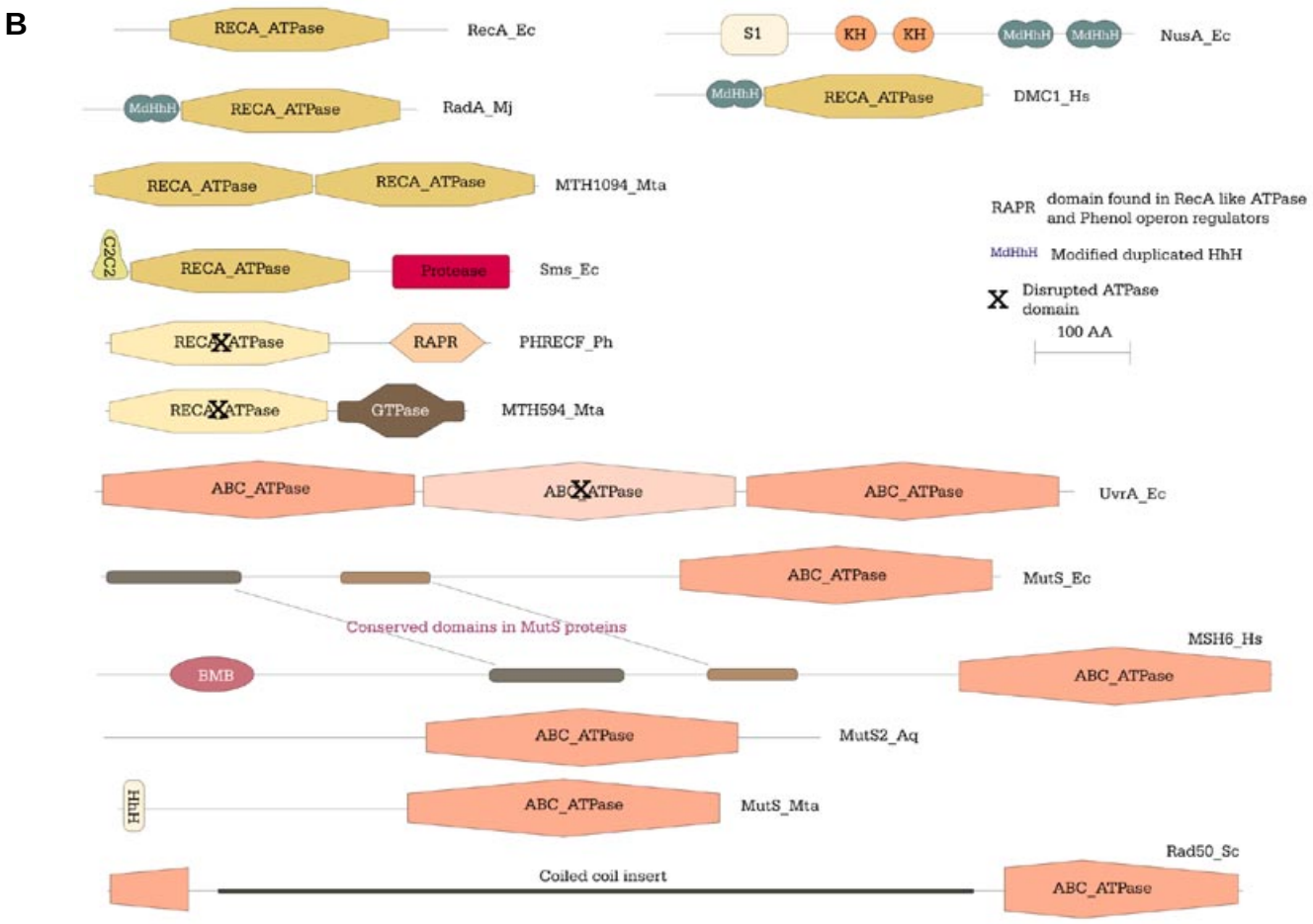
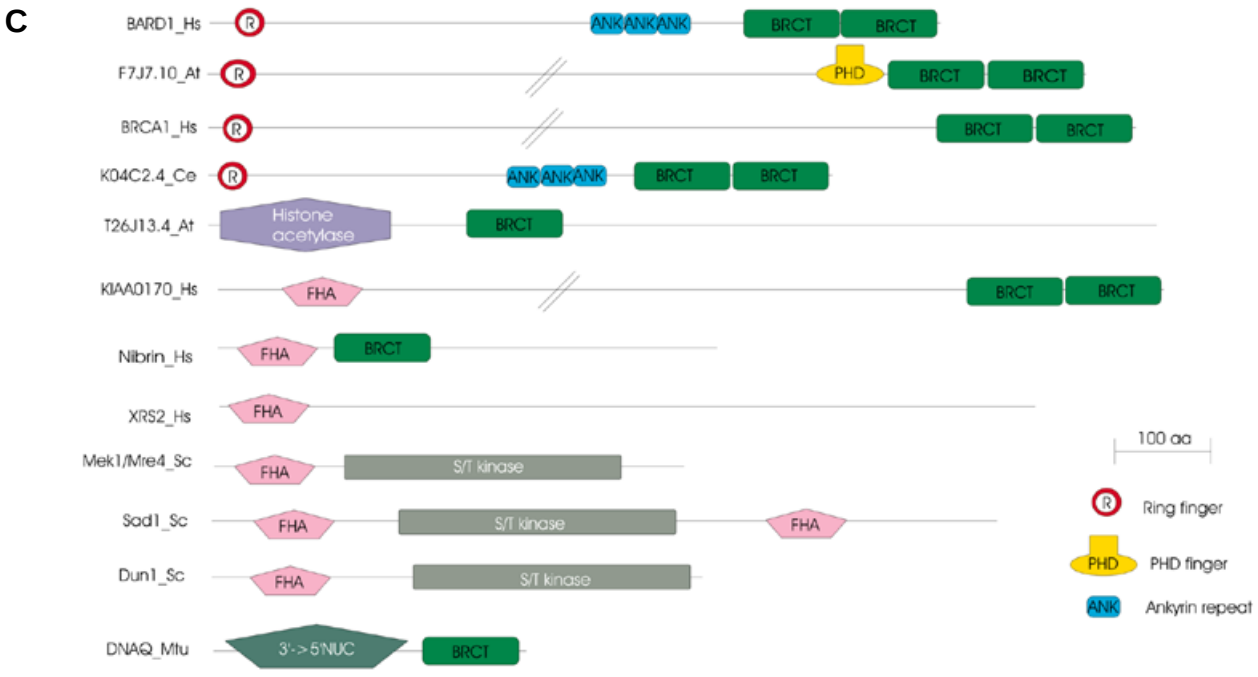


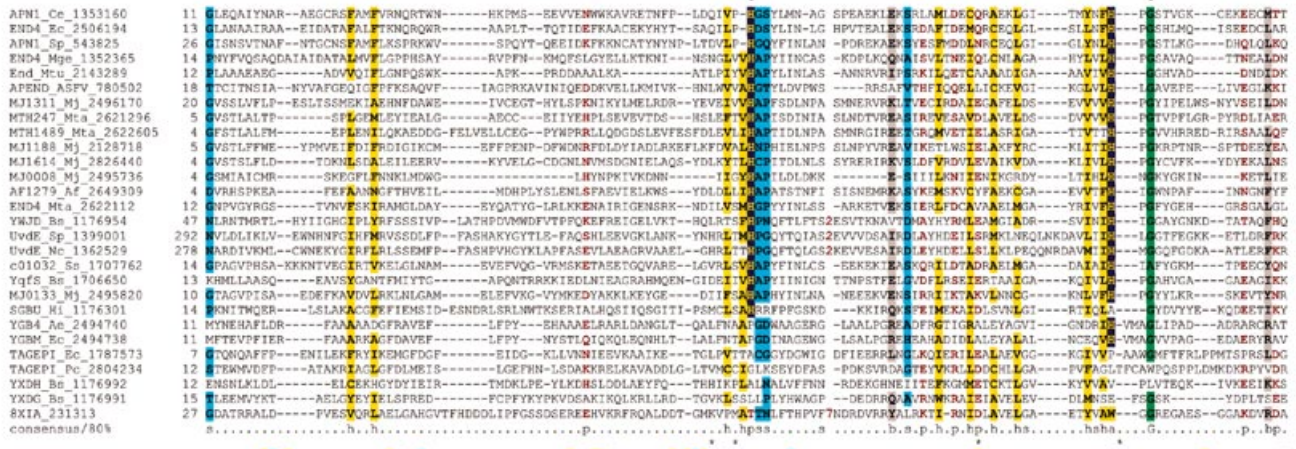
Figure 1. (Above and opposite). Domain architectures of selected repair proteins. (**A**) Helicases and nucleases; (**B**) ATPases; (**C**) BRCT domain-containing proteins. The figure is approximately to scale. Crossed symbols indicate domains with disrupted functional motifs. ERCC1 contains an apparently inactivated nuclease domain, whereas Mfd and ERCC4 contain inactivated helicase domains; ERCC4 also contains disrupted HhH motifs. SF1/2 stand for superfamily I and II helicases, respectively; pol, DNA polymerase I catalytic domain; URI, UvrC, intron nuclease domain; UVRBC, a possible adaptor domain shared by UvrB and UvrC; ENDOV, a nuclease domain found in endonuclease V and UVRC; FCL, Fe cluster; HhH, helix–hairpin–helix nucleic acid-binding domain; MdHhH, modified HhH domain; HRD, a predicted nucleic acid binding domain found in some recQ family helicases; RqC, RecQ C-terminal (domain); C2C2 (‘little finger’), predicted small, metal-binding DNA binding module; PHD, a distinct type of Zn finger; ANK, ankyrin repeat; OB, oligonucleotide-binding domain; Nuc, nuclease; BRCT, BRCA1 C-terminal (domain); FHA, forkhead homology-associated (domain); S/T kinase, serine/threonine kinase. The different shapes of nuclease domains indicate the different nuclease (super)families described in the text and Figure 2. In (B) the NusA protein (a transcription factor not involved in repair) is shown to illustrate the conservation of the modified HhH domain that is also found in eukaryotic and archaeal RecA orthologs; NusA contains additional nucleic acid-binding domains, namely KH and S1. Other designations are directly on the figures. Double slash (//) shows that a middle portion was omitted in some large proteins. The proteins are identified by their gene names or names from the SWISS-PROT database, and the source species is indicated after an underline. The species abbreviations are: Af, *A.fulgidus*; Aq, *A.aeolicus*; Ce, *C.elegans*; Dm, *Drosophila melanogaster*; Ec, *E.coli*; Hs, *Homo sapiens*; Mj, *M.jannaschii*; Mta, *M.thermoautotrophicum*; Mtu, *M.tuberculosis*; Ph, *P.horikoshii*; Sc, *S.cerevisiae*.

proteins that interact with DNA contain distinct, compact DNA-binding domains that combine with different enzymatic or adaptor domains (Fig. 1A–C). Typically, DNA-binding domains show much less sequence conservation than enzymes and, in many cases, recognition of these domains requires careful application of sensitive computer methods. Particularly striking is the case of the nucleic acid-binding HhH module that appears to be the most common DNA-binding domain in repair systems but has been identified only recently by a combination of experimental and computer analysis (45,53). In the course of the present study, we identified previously undetected, distinct versions of the HhH domain in three families of repair proteins, namely the ERCC4 nuclease/helicase family, the archaeal and eukaryotic members of the RadA/RecA family and the UmuC/REV1 superfamily (Fig. 1A and B; data not shown). In each case, the amino acid patterns typical of HhH domains are modified and not easily recognizable, but show a relationship to similarly modified HhH domains seen at the C-terminus of the RNA-binding protein NusA from certain bacteria (e.g. *E.coli* and *Chlamydia*). Statistically significant similarity to classical HhH domains, in particular those in the DNA ligases, can be demonstrated for these domains only in iterative database searches. A completely different type of a DNA-binding domain is represented by iron-binding cysteine clusters (53) that are conserved in a subset of the RecB nuclease family and the endonuclease III family (Figs 1A and 2E).

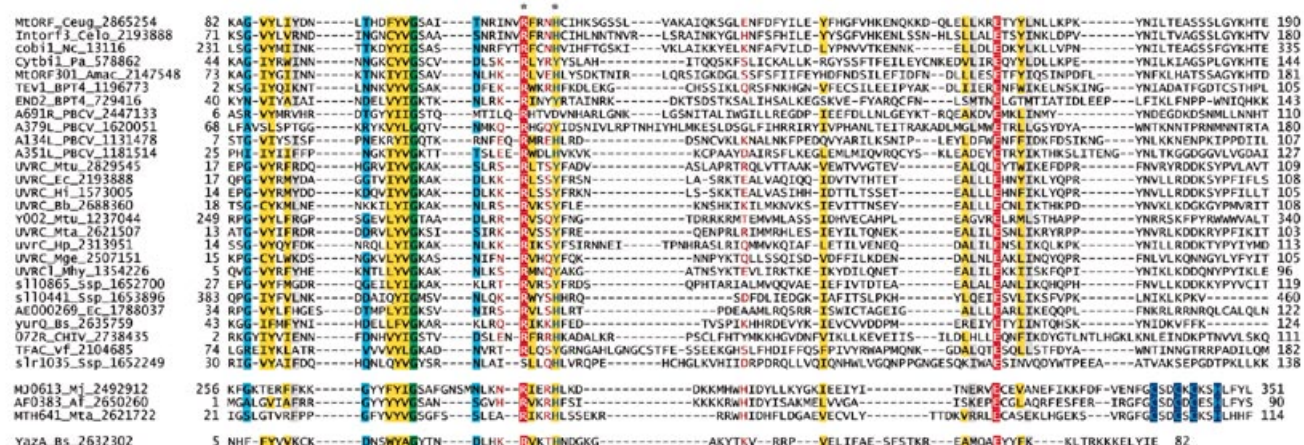

B
RECA_ATPase
RecA_Ec
S1
KH
KH
MdHhH
MdHhH
NusA_Ec
MdHhH
RECA_ATPase
RadA_Mj
MdHhH
RECA_ATPase
DMC1_Hs
RECA_ATPase
RECA_ATPase
MTH1094_Mta
RAPR domain found in RecA like ATPase and Phenol operon regulators
MdHhH Modified duplicated HhH
X Disrupted ATPase domain
100 AA
C2C2
RECA_ATPase
Protease
Sms_Ec
RECXATPase
RAPR
PHRECF_Ph
RECXATPase
GTPase
MTH594_Mta
ABC_ATPase
ABCXATPase
ABC_ATPase
UvrA_Ec
ABC_ATPase
MutS_Ec
Conserved domains in MutS proteins
MSH6_Hs
BMB
ABC_ATPase
ABC_ATPase
MutS2_Aq
HhH
ABC_ATPase
MutS_Mta
Rad50_Sc
Coiled coil insert
ABC_ATPase


C
BARD1_Hs
R
ANK ANK ANK
BRCT
BRCT
F7J7.10_At
R
PHD
BRCT
BRCT
BRCA1_Hs
R
BRCT
BRCT
K04C2.4_Ce
R
ANK ANK ANK
BRCT
BRCT
T26J13.4_At
Histone acetylase
BRCT
KIAA0170_Hs
FHA
BRCT
BRCT
Nibrin_Hs
FHA
BRCT
XRS2_Hs
FHA
Mek1/Mre4_Sc
FHA
S/T kinase
Sad1_Sc
FHA
S/T kinase
FHA
Dun1_Sc
FHA
S/T kinase
DNAQ_Mtu
3'->5'NUC
BRCT
100 aa
R Ring finger
PHD PHD finger
ANK Ankyrin repeat

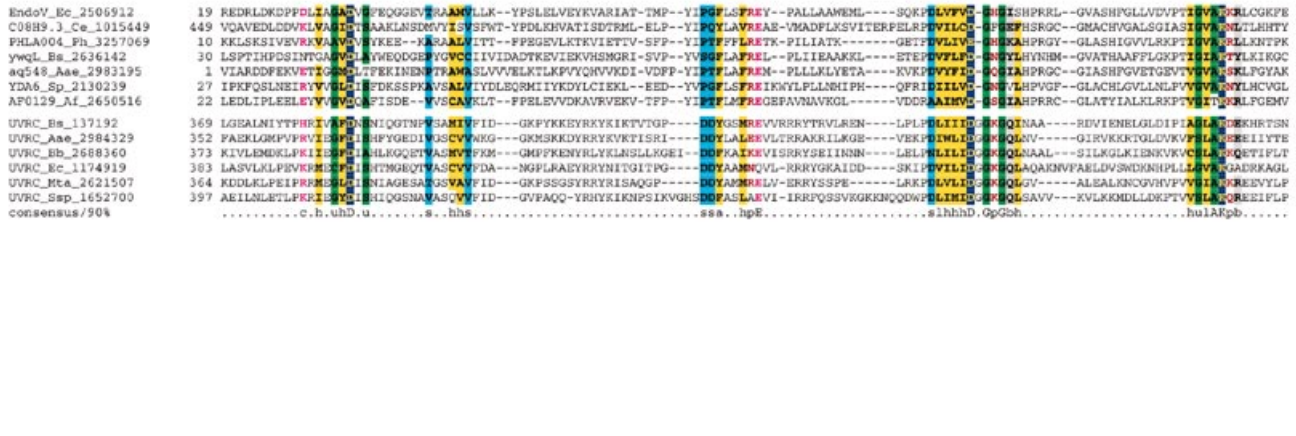
A



B



C



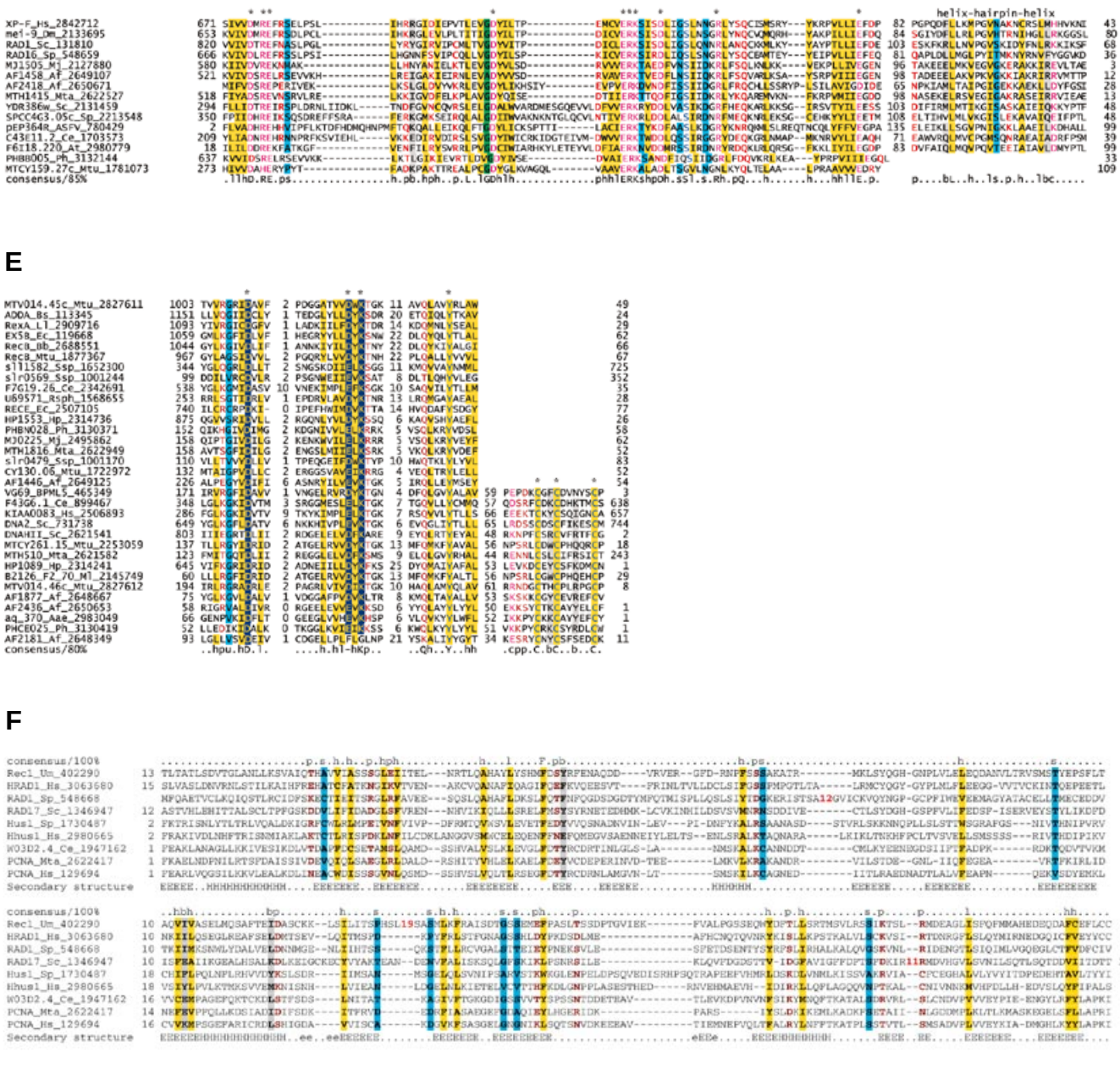


Figure 2. (Above and opposite). Multiple sequence alignment of previously undetected and expanded domain families of repair proteins. (**A**) AP endonuclease/ENDO4 superfamily; (**B**) Uri domain endonuclease family; (**C**) EndoV endonuclease family; (**D**) RAD1/ERCC4 endonuclease superfamily; (**E**) RecB nuclease domain family; (**F**) PCNA family. The alignments were constructed on the basis of the PSI-BLAST results using the ClustalW program. The left column includes the protein names from the SWISS-PROT database or gene names, and the Gene Identification (GI) numbers (after the underscore). The species abbreviations are: ASFV, African Swine Fever Virus; BPML5, *M.leprae* bacteriophage 5; BPT4, bacteriophage T4; CHIV, Chilo Iridiscent virus; NPV, Nuclear Polyhedrosis virus; PBCV, *Paramecium bursaria* Chlorella virus; Aa, *A.aeolicus*; Aae, *Alcaligenes eutrophus*; Af, *A.fulgidus*; Amac, *Allomyces macrogynus*; At, *Arabidopsis thaliana*; Bb, *Borrelia burgdorferi*; Bs, *B.subtilis*, Ce, *C.elegans*; Celo, *Chlorogonium elongatum*; Ceug, *Chlamydomonas eugametos*; Dm, *D.melanogaster*; Hs, *H.sapiens*; Ct, *Chlamydia trachomatis*; Ec, *E.coli*; Hi, *H.influenzae*; Hp, *Helicobacter pylori*; Ll, *Lactococcus lactis*; Mj, *M.jannaschii*; Mge, *Mycoplasma genitalium*; Mhy, *Mycoplasma hyorhinis*; Mpn, *Mycoplasma pneumoniae*; Mta, *M.thermoautotrophicum*; Mtu, *M.tuberculosis*; Mpn, *Mycoplasma pneumoniae*; Nc, *Neurospora crassa*; Ngo, *Neisseria gonorrhoeae*; Pa, *Podospora anserina*; Pf, *Pyrococcus furiosus*; Ph, *P.horikoshii*; Pv, *Phaseolus vulgaris*; Rsph, *Rhodopseudomonas spheroides*; Sag, *Streptococcus agalactiae*; Sc, *S.cerevisiae*; Sp, *S.pombe*; Ss, *Synechocystis* sp.; St, *Streptococcus thermophilus*; Tp, *T.pallidum*; Um, *Ustilago maydis*; Vf, *Vicia faba*. In each panel, a consensus derived using the indicated percentage cut-off is shown, and the respective alignment columns are highlighted through differential coloring; **b** indicates a 'big' residue (E,K,R,I,L,M,F,Y,W), **h** indicates hydrophobic residues (A,C,F,I,L,M,V,W,Y), **s** indicates small residues (A,C,S,T,D,N,V,G,P), u indicates 'tiny' residues (G,A,S), **p** indicates polar residues (D,E,H,K,N,Q,R,S,T), c indicates charged residues (K,R,D,E,H), and '–' indicates negatively charged residues (D,E). The conserved charged residues that may be directly involved in enzymatic catalysis are indicated by asterisks. The distances from the aligned regions to the protein termini and the distances between the conserved blocks, where more variable regions were omitted, are indicated by numbers. In (F), the secondary structure elements derived from the crystal structure of PCNA are shown underneath the alignment; E indicates extended conformation (β-strand), and H indicates α-helix.

Some conserved domains in repair proteins are implicated in DNA binding even in the absence of direct experimental characterization for any representative, primarily on the basis of their predicted compact structure, small size and absence of conserved polar residues that could be involved in a catalytic activity. An example of such predicted nucleic acid-binding domain awaiting experimental corroboration is the HRD domain found in a subset of the RecQ family helicases, e.g. human Werner's and Bloom's syndrome gene products, and in RNase D (54).

Adaptor domains

The components of the repair machinery typically function in the form of macromolecular complexes that consist of multiple, diverse subunits. Therefore, in addition to DNA-binding domains, adaptor domains, that is domains that mediate protein–protein interactions between the components of repair complexes as well as between repair proteins and other cellular components, have a prominent role in repair. Adaptor domains are particularly important in eukaryotes where repair is intimately connected to the dynamics of chromatin-associated protein complexes and their alteration linked to the progression of the cell cycle, but prokaryotic adaptors also seem to exist. An example of likely bacterial adaptors is the domain shared by the UvrB (C-terminal domain) and UvrC proteins and implicated in the formation of the complex between these proteins (Fig. 1A; 55).

Arguably, the most important adaptor domain involved in eukaryotic repair is the BRCT (BRca1 C-terminal) domain that has been detected in a vast variety of proteins involved in repair and cell cycle checkpoint regulation and may provide the critical connections between these processes (56,57; see also the discussion below). The BRCT domain occurs on its own in multiple copies as in yeast RAD9 or combines with a variety of enzymatic and DNA-binding domains as in terminal nucleotidyl transferases (TdT), REV1 and DNA ligases. In those instances where the function of the BRCT domain has been determined experimentally, BRCT domains of different repair proteins, such as DNA ligases III, XRCC1, poly(ADP-ribose) polymerase (PARP) and BRCA1, appear to mediate specific protein–protein interactions (58–60), which provides for the formation of protein complexes involved both in repair and in cell cycle checkpoints.

Examination of the protein sequences that have become available subsequent to the previous analyses of the BRCT domain revealed several interesting new occurrences (Fig. 1C). Specifically, and unexpectedly, we found that an uncharacterized plant protein not only is highly similar to mammalian BRCA1 and BARD1 but also mimics their unique domain organization in terms of the relative location of the BRCT and RING domains (Fig. 1C). The plant counterpart, however, contains an additional domain, namely a PHD finger, which suggests DNA binding. Furthermore, we showed that the trypanosomal protein with similarity to the BRCT domain that was suspected to be a false positive (12) contains a bona fide copy of the domain, thus expanding the BRCT domain distribution outside the crown group of the eukaryotes. Another novel domain architecture was observed in a protein from *M.tuberculosis* that combines a 3′–5′ exonuclease domain with a C-terminal BRCT domain (Fig. 1C). This is the first combination of a BRCT domain with an enzymatic domain other than DNA ligase in a bacterium.

The list of adaptor domains involved in repair and its interaction with cell cycle checkpoints is growing. The FHA (forkhead homology associated) domain has been detected in a variety of proteins with diverse functions, including protein kinases implicated in DNA damage response (61) and Xrs2 which participates in the repair of double strand breaks (62). The recent demonstration that the FHA domain of the RAD53 kinase interacts with the phosphorylated form of the BRCT protein RAD9 (63) indicates that FHA is a repair-checkpoint adaptor that may recognize phosphorylated proteins, perhaps even specifically phosphorylated BRCT domains. This possibility is of particular interest given the independent evolution of proteins combining the FHA and BRCT domains on at least two occasions (Fig. 1C).

The recently described HORMA domain that has been detected in the yeast REV7 protein involved in translesion DNA synthesis and in proteins that participate in the spindle assembly checkpoint and synaptonemal complex formation in meiosis, such as MAD2 and HOP1, is an example of an adaptor with a more limited distribution which, however, may have a critical role in linking repair with the cell cycle (64).

A protein with versatile adaptor functions is the proliferating cell nucleus antigen (PCNA) that originally has been identified as the sliding clamp that is required to increase the eukaryotic DNA polymerase processivity (65). More recently, it has been shown that PCNA is required for NER and MMR and interacts with a variety of repair proteins (65,66). In the course of the present analysis, we showed that PCNA is homologous to a group of proteins involved in repair and DNA damage checkpoints that include yeast RAD17, *S.pombe* Rad1 and Hus1, REC1 from *Ustilago*, and their mammalian orthologs (Fig. 2F). The similarity between PCNA and the repair proteins is subtle but statistically significant; for example, a PSI-BLAST search initiated with the sequence of the *Methanobacterium autotrophicum* PCNA ortholog retrieved the *S.pombe* Rad1 sequence with an e-value of 0.003 on the second iteration, with the rest of the homologous repair proteins detected on the subsequent iterations. The alignment spans the entire length of PCNA, and the observed conserved motifs are compatible with the PCNA 3D structure (Fig. 2F), supporting the notion that these proteins have the PCNA fold (67). Two of these proteins, namely the *Ustilago* REC1 and the human ortholog of Rad1, have been shown to possess nuclease activity (68,69). PCNA is highly conserved amidst the eukaryotes and is homologous to the bacterial DNA pol III β subunits (67,70). None of these well studied proteins has been shown to possess any nuclease activity, suggesting that this property may have been secondarily derived in the Rad1 subfamily of the family of PCNA-related proteins. It seems possible, on the other hand, regardless of the nuclease activity, that at least some of these proteins bind DNA and play a role in the assembly of repair-specific complexes. The yeast RAD24 and the Rad17 protein from *S.pombe*, which function in the same checkpoint with yeast RAD17 and *S.pombe* Rad1 and hus1, respectively (71), are homologs of the clamp loader ATPases involved in replication and may facilitate the formation of such complexes in an ATP-dependent fashion. The determinants of protein–protein interactions in PCNA have been mapped to loops (66,72) that are not highly conserved in the repair proteins which suggests that the actual partners of these proteins may be different from those of PCNA.

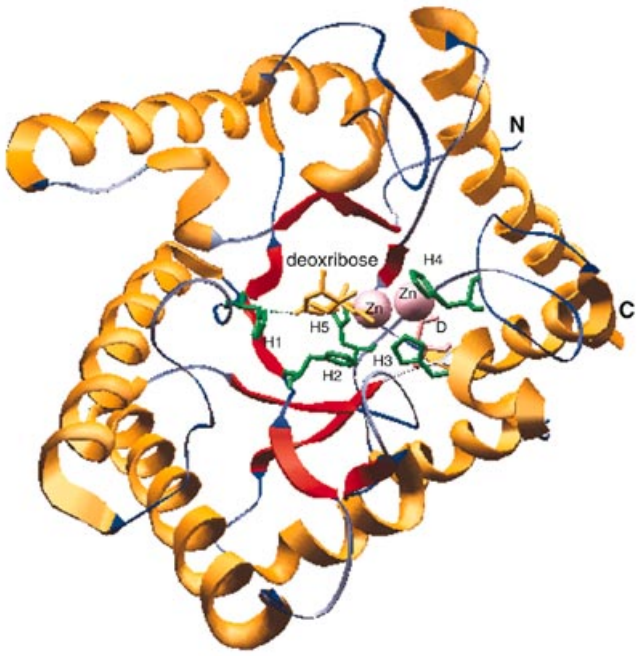


Figure 3. A structural model of *E.coli* endonuclease IV built using the xylose isomerase structure as a template. The structural manipulations were done using the SWISSPDBviewer program. Using the multiple alignment shown in Figure 1A, a composite target sequence of the AP endonuclease was constructed, with the xylose isomerase structure (PDB coded 8XIA) serving as a template. The alignment of this composite sequence with 8XIA was further adjusted so that the energy of the target was globally minimized using a Sippl-like field. The resulting refined alignment was submitted as a PROMODII job, and the model was obtained. The deoxyribose of DNA appears to be positioned in the endonuclease model exactly as the xylose molecule is in xylose isomerase. The strands are colored red, the helices gold, the conserved aspartate orange, the conserved histidines (labeled H1–H5 from the N- to the C-terminus) green.

PHYLETIC DISTRIBUTION AND EVOLUTION OF REPAIR SYSTEMS

The biochemical studies on repair systems have been mostly limited to a few model species, such as *E.coli*, the yeast *S.cerevisiae*, and humans. Therefore, analysis of the distribution of orthologs of repair proteins from these organisms in different phylogenetic lineages not only provides the material for evolutionary scenarios but effectively, amounts to the reconstruction of the repair systems in poorly studied organisms. Evidently, the completeness and precision of such a reconstruction depends both on the quality of analysis and on the level of conservation of the repair mechanisms between the organisms in question and one of the model species.

The most striking aspect of the phyletic distribution of repair systems that becomes apparent through the comparison of complete protein sets from distant species is that while the repertoire of principal domains involved in repair, such as several distinct types of helicases and nucleases, is to a large extent conserved in all cells, the number of orthologous or even clearly functionally equivalent repair proteins that are shared by all the three superkingdoms is very small. By contrast, there is a much greater number of repair proteins that are conserved in one or two superkingdoms (Tables 1 and 2).

REPAIR PROTEINS CONSERVED IN ALL THREE SUPERKINGDOMS OF LIFE

There seem to be no known repair proteins with an identical domain arrangement conserved in bacteria, archaea and eukaryotes. There are, however, a few highly conserved proteins with limited variations of domain architecture, of which the only one encoded in all genomes sequenced so far and apparently truly universal, is the RecA/RadA recombinase, which plays a central role in DNA recombination and RER (73,74). While RecA(RadA) appears to have been vertically transmitted throughout the history of life, its evolution has been accompanied by notable variations on the main theme, the most important being the fusion with a modified HhH domain that is shared by archaea and eukaryotes (Fig. 1B, and above). The presence of an additional domain predicted to bind single-stranded DNA in the archaeal and eukaryotic RadA proteins suggests differences in the mode of their interaction with DNA, compared to bacterial RecA proteins. Duplications of the RecA ATPase domain accompanied by domain accretion and divergence seem to have occurred independently in different phylogenetic lineages (Fig. 1B). An apparent early series of events in bacterial evolution produced the *sms* gene coding for a protein involved in radioresistance (75,76) and containing a RecA domain flanked by a C2C2 Zn finger domain and a predicted serine protease domain that may be inactivated in some species (Fig. 1B; 77). In archaea, additional intramolecular duplications and fusions of the RecA family ATPase domains are observed, and in some of these proteins, the conserved motifs in the ATPase domain are disrupted, suggesting its inactivation (Fig. 1B); some of these proteins may have been recruited for roles in processes other than repair.

Another universally conserved domain that is found, however, in significantly different structural and functional contexts in bacteria, on one hand, and in archaea and eukaryotes, on the other hand, is the FLAP nuclease (78–80). In archaea and eukaryotes, these nucleases (e.g. yeast RAD2 and RAD27) cleave recombination and repair intermediates containing overlapping 5′-flaps at sites of nicks; they also possess 5′–3′ exonuclease activity that may be involved in the hydrolysis of these flaps (78,79). The bacterial ortholog of the FLAP endonucleases is the N-terminal, 5′–3′ exonuclease domain of DNA polymerase I (Fig. 1A) that is involved in the excision of damaged single-stranded DNA fragments at nick sites (81). In two groups of bacteria, namely *Mycoplasma* and *Aquifex*, the 5′–3′ exonuclease domain is encoded by a separate gene. Both polymerase-associated and stand-alone bacterial exonucleases share the HhH domain, emphasizing the orthologous relationship with the archaeal and eukaryotic FLAP nucleases. Iterative searches identify several novel members of this family in eukaryotes and bacteria (e.g. *Drosophila* Asteroid), some of which may be as yet unknown repair proteins. This example clearly illustrates the distinct evolutionary histories of the repair systems in the three superkingdoms, even when well conserved, universal domains are involved.

Several other repair proteins, though not ubiquitous, are found in most representatives of all three superkingdoms (Table 1). The most striking example of this kind are the SMC-like ATPases and the associated nucleases. These ATPases (typified by the *E.coli* SbcC protein) belong to the ABC superfamily but have an inserted large coiled-coil domain between the P-loop and the Mg^{2+}-binding motif that together comprise the ATP-binding site. They are seen in almost all complete genomes (Table 1), and in eukaryotes, are

involved in ATP-dependent, large-scale modifications of the chromatin structure (82,83). The SMC-like ATPases form complexes with the equally common nucleases of the calcineurin-like phosphoesterase superfamily, such as bacterial SbcD-like proteins and eukaryotic Mre11-like proteins (84–86). It seems likely that this ATPase-nuclease pair was vertically inherited in all life forms with a loss in a few lineages.

Other conserved repair proteins found in all three superkingdoms, with a varying degree of representation among specific lineages, include photolyases (*phrB* gene product in *E.coli*), endonuclease III (*nth* and *mutY*), exonuclease III (*xthA*), 8-oxo-dGTPase (*mutT*) and the UmuC protein superfamily. Each of these enzymes is involved in a basic repair function (1 and references therein), but their activities are, in principle, dispensable as each of them is missing in some of the bacterial or archaeal species with small genomes (Table 1).

REPAIR PROTEINS AND PATHWAYS CONFINED TO ONLY ONE OR TWO OF THE SUPERKINGDOMS

The protein families discussed in the previous section represent the relatively small number of cases when homologous domains arranged in similar, if not identical, combinations appear to perform similar functions in repair in all three superkingdoms. By contrast, most of the repair systems have more limited phyletic distribution, which in some instances may suggest plausible scenarios for their evolution.

Repair systems of bacterial origin

Several repair systems are essentially unique to bacteria but some of these additionally are seen in eukaryotes, to the exclusion of the archaea (Table 1), which may suggest horizontal gene transfer, in most cases probably from the mitochondrial genome to the eukaryotic nuclear genome. The UvrABC excisionase, together with the UvrD helicase that is functionally coupled to it, are the principal components of NER in bacteria (4) and are encoded in all bacterial genomes sequenced to date, including the minimal genomes of *Mycoplasma*. Outside the bacteria, however, this system has been detected in only one archaeon, namely *Methanobacterium thermoautotrophicum*. *Methanobacterium thermoautotrophicum* has a complete operon including the *uvrA*, *B* and *C* genes, and UvrD encoded elsewhere in the genome, which strongly suggests horizontal transfer from bacteria. The domain architecture of all three excisionase subunits is conserved throughout bacteria, but the presence of the Uri and EndoV nuclease domains in other contexts (Fig. 1A) suggests that these nucleases had been repeatedly recruited for distinct functions, which may include other repair systems.

The second widespread bacterial repair system is the RuvAB(C) complex, which is the Holliday junction resolvase and the key component of bacterial RER (87,88). Interestingly, RuvC, the endonuclease subunit, is not detectable in *Mycoplasma* and spirochaetes, suggesting that a distinct nuclease may have been recruited in these bacteria for the participation in Holliday junction resolution. As in the case of the UvrABCD system, each of the Ruv proteins contains well known ancient conserved domains (Table 1) but orthologs of these proteins so far have been detected only in bacteria.

A different phylogenetic pattern was observed among the components of the base MMR system (5,89). This system depends primarily on two proteins containing ATPase domains of different structures, namely MutL (90,91) and MutS (28), both of which are highly conserved among bacteria, though missing in *Mycoplasma*. Only the MutS family proteins are seen in the archaea *M.thermoautotrophicum* (with an additional HhH domain) and *Pyrococcus horikoshi*. This finding is of particular interest as these are so far the only genomes in which a gene for MutS is not accompanied by a MutL gene, suggesting the possibility of functional uncoupling between these MMR system components.

Phylogenetic analysis of the MutS protein sequences shows that a gene duplication resulting in two distinct forms of MutS had occurred very early in bacterial evolution (data not shown). This is supported, in particular, by the presence of both forms in bacteria from several major lineages, such as *Aquifex aeolicus*, *B.subtilis* and *Synechocystis*. There is a major expansion of genes encoding MutL and MutS homologs in eukaryotes, with at least five or six members found in each eukaryotic genome. This expansion apparently involves functional diversification, in particular between nuclear and mitochondrial DNA repair. In the course of this analysis, we observed that one of the families of eukaryotic MutS homologs (GMBP1) contains an additional domain (BMB domain in Fig. 1A), which is also found in eukaryotic chromatin-associated proteins, such as BS69 and BR140 (L.Aravind, unpublished), and may link the eukaryotic MMR system with the chromatin. The most likely scenario for the evolution of the MMR system involves gene transfer from mitochondria to the eukaryotic nuclear genome, with subsequent multiple duplications. This scheme is compatible with the role of some of the eukaryotic MutL and MutS homologs in mitochondrial repair (92) and with the topology of phylogenetic trees (data not shown).

Illegitimate recombination in bacteria and eukaryotes is suppressed by the RecQ helicase family members, which accordingly appear to play a major role in the maintenance of chromosomal integrity (93,94). There are two highly conserved RecQ paralogs, which differ by the presence or absence of the putative DNA-binding HRD domain (54); one or both paralogs may be present in the same genome amidst different bacterial lineages. Multiple orthologs of both of these RecQ-like helicases are detectable in eukaryotes but not in archaea. Remarkably, two human gene that are mutated in hereditary diseases associated with repair defects, namely Bloom's and Werner's syndromes (95,96), encode HRD domain-containing helicases of the RecQ family (Fig. 1A). The evolutionary history of the RecQ family of helicases appears be analogous to that of the MMR system and probably included horizontal gene transfer from mitochondria to the eukaryotic nuclear genome.

The only repair protein that is conserved in most bacteria and apparently all archaea, to the exclusion of eukaryotes, is the RecJ 5′–3′ exonuclease, which belongs to the recently identified 'DHH' superfamily of phosphohydrolases (97). The eukaryotic members of this superfamily (e.g. the *Drosophila* Prune protein) are only distantly related to RecJ and do not seem to be involved in repair. RecJ has been implicated both in RER and in the post-incision removal of 5′-deoxyribose phosphate in BER (98,99) but it appears that the common function of this nuclease underlying its notable conservation in bacteria and archaea remains to be identified.

Additional, specifically bacterial repair pathways rely on distinct members of the ABC superfamily of ATPases, such as RecN and RecF, helicases, e.g. RecG (100) and accessory, single-stranded DNA-binding proteins, such as RecO and RecR

(101). The evolution of RecR is of particular interest as it is a clear case of recruitment of an enzymatic domain, namely the recently identified common catalytic domain of DNA primases and topoisomerases (Toprim domain; 102), for a non-enzymatic function.

Bacteria have evolved a unique regulatory system, which allows them to produce a complex response to DNA damage. This system depends on the DNA-binding transcription regulators LexA (103) and UmuD (104) containing a C-terminal signal peptidase-like domain, which catalyzes RecA-dependent autoproteolysis of these proteins, thus activating the DNA-binding domain. LexA is a general transcriptional regulator of repair functions; LexA orthologs are limited in their distribution to several bacterial lineages. The theme of the association of proteolysis with repair, however, appears to be more general. The bacteria-specific repair ATPase Sms consists of three domains (Fig. 1B), one of which is a protease domain of the Lon superfamily of serine proteases (predicted to be active in some bacteria but apparently inactivated in others). The function of this protease in repair, which conceivably may involve an as yet uncharacterized cleavage of specific proteins with a regulatory effect, remains to be clarified.

Coupling of transcription and repair appears to confer a definite selective advantage as it enables the organism to repair functional genes as they are expressed and thus escape the immediate effects of deleterious mutations resulting in non-functional proteins. This coupling seems to have evolved independently in bacteria and in eukaryotes. The bacterial version is dependent on the superfamily II helicase Mfd/TRCF (105,106) that is conserved in several bacterial lineages and contains a second, apparently inactivated helicase domain whose function could be the recruitment of other repair proteins (Fig. 1A and Table 1).

Several other repair pathways are restricted to just a few groups of bacteria (Table 1); a thoroughly studied example is the RecBCD helicase–exonuclease complex, which is the central component of RER. In some cases, recruitment of a repair enzyme in a subset of bacteria from rather unexpected sources seems likely. Thus the dcm and dam methylases (107) appear to have been recruited from restriction system methylases of phage origin. Similarly, the MutH endonuclease involved in MMR and so far found only in *E.coli* and *H.influenzae* probably has been derived from a restriction endonuclease related to Sau3 (108).

Repair systems of archaeal and eukaryotic origin

The NER system, transcription-repair coupling components and the vast repertoire of regulatory proteins distinguish the eukaryotic repair systems from bacterial ones. While the NER system includes components that individually trace back to the common ancestor of the archaea and eukaryotes, the transcription-repair coupling mechanism and the regulatory apparatus seem to be true eukaryotic inventions that probably have evolved in response to the diversification of the eukaryotic chromatin structure and cell cycle control. Even within the eukaryotes, while the core machinery appears to be conserved throughout, there are several notable, lineage-specific modifications of the regulatory system.

The understanding of the core eukaryotic repair systems has largely been derived from the RAD complementation groups in yeast (109) and the Xeroderma pigmentosum complementation groups in humans (110) (Table 2). The intersection of the results produced by these principal lines of research delineates the conserved central components of eukaryotic NER. The eukaryotic NER system is built up of a number of distinct helicases and nucleases. The helicases include ERCC2 (Xp-D) (111), ERCC3 (Xp-B) (112) and ERCC6 (Cs-B) (113). The ERCC2 helicase is conserved in all eukaryotes sampled so far and shows a distant but apparently orthologous relationship with the DinG helicase (114) seen in several bacteria and the archaeon *M.jannaschii*, suggesting an ancient involvement in repair. However, beyond the general helicase role, the members of this family appear to have undergone functional differentiation following independent duplication in different phylogenetic lineages. For example, the eukaryotic CHL1 helicase, a member of the ERCC2 family, has a role in maintaining the chromatin integrity (115).

The ERCC3 helicase family shows an unusual phyletic distribution—in addition to its conservation in eukaryotes, it is also present in the archaeon *A.fulgidus*, the bacteria *Mycobacterium leprae* (116) and *Treponema pallidum*, African swine fever virus and some bacteriophages, suggesting multiple horizontal gene transfer events. Given the lack of orthologs of other members of the eukaryotic-type NER complex in bacteria and archaea, it is unlikely that these scattered ERCC3 orthologs share functional details with the eukaryotic enzyme.

The ERCC6 helicase belongs to the ancient SWI/SNF family that is conserved in bacteria and eukaryotes. In eukaryotes, however, this family has undergone a striking expansion, with 17 paralogous members in yeast (117), many of which are involved in repair. Bacterial helicases of the HepA family, which are orthologous to the ERCC6 family (118), may be involved in repair and specifically in the repair–transcription coupling (119), but this family is represented by only one or two members in each bacterial genome when present. Thus it is obvious that the SWI/SNF family has attained its current functional differentiation only after the origin of the eukaryotes. This must have been an early event in eukaryotic evolution since for a number of these helicases, orthologous relationships can be traced in yeast, plants and animals. In some of these orthologous sets, such as RAD5 (120) and RAD16 (121), a unique domain organization, with a RING finger inserted into the helicase domain, between the helicase motifs 5 and 6 (Fig. 1A), is conserved throughout the Eukarya. This domain architecture probably had evolved early in eukaryotic evolution as a device for tethering the helicase to chromatin.

The nuclease components of the NER system also are highly conserved, and as noted above, ERCC4 is seen in archaea as well, fused to an apparently active N-terminal helicase domain. The other nucleases in this pathway, such as Xp-G, Rad2 and Rad27, are members of the universally-conserved FLAP/FEN family (122). Another NER component is the UV-damaged DNA-binding protein (UV-DDB) which partially complements the XP-E defect (123). UV-DDB is a member of a family that has two additional paralogs conserved in eukaryotes, one of which is a component of the polyA cleavage specificity factor (CPSF-A) (Table 2). In this context it is interesting to note that another repair protein SNM1, which is involved in UV cross-link repair in yeast (124), is homologous to other CPSF subunits that contain a metallo-β-lactamase domain (125).

The regulation of repair and its connection with cell cycle checkpoints are the most dramatic distinguishing features of the eukaryotic repair system that have undergone considerable evolution after the divergence of the eukaryotes from the other superkingdoms of life. The proteins providing for these features typically have no orthologs in bacteria or archaea, even though

some of the adaptor domains are conserved. The understanding of the likely structural basis of the repair-checkpoint coupling has been significantly advanced through the discovery of a single domain—the BRCT domain that appears to be the most common adaptor in the eukaryotic repair machinery. The yeast genome encodes 10 BRCT-containing proteins (57), and the number of these proteins encoded in the genomes of multicellular eukaryotes is expected to be even greater. As discussed above, certain distinct domain architectures of BRCT-containing proteins are highly conserved in evolution. Generally, however, domain shuffling seems to be the predominant trend in the evolution of the BRCT-containing proteins. Thus, of the 10 yeast BRCT-containing proteins, only three, namely the DNA ligase, DNA polymerase subunit 2 (DPB11) and the REV1 nucleotidyltransferase, are represented by orthologs with a conserved domain arrangement in *Caenorhabditis elegans*. Conversely, *C.elegans* encodes a number of BRCT-containing proteins with unique domain architectures.

The BRCT domain thus far has not been detected in archaea but is invariably present at the C-terminus of bacterial DNA ligases. This phyletic distribution suggests that similarly to several other components of the repair system (e.g. MMR components), the BRCT domain most likely had invaded the eukaryotic genomes by gene transfer from bacteria and had subsequently undergone a dramatic expansion in the eukaryotes. The detection of a BRCT domain protein in trypanosomes indicates that the proposed horizontal gene transfer event dates to a very early stage in the evolution of eukaryotes.

There are other proteins with very diverse functions that appear to connect the eukaryotic repair systems with chromatin. Typically, such proteins contain eukaryote-specific adaptor domains, such as the RING finger (126) in some of the SWI family helicases and other proteins like RAD18, the WD40 repeats in CS-A (127), and ubiquitin and duplicated ubiquitin hydrolase domains in Xp-C/Rad23 (128). The signal transmission from damaged DNA to the checkpoint machinery relies upon a phosphorylation cascade that includes FHA domain-containing kinases, such as SAD1 (129) and DUN1 (130), and the ATM kinases (131) of the lipid kinase superfamily. Finally, several eukaryotic proteins regulate the repair machinery at the level of transcription; the best characterized representatives of this group are p53 and retinoblastoma (Rb) (132). These regulators appear to have evolved in specific groups of eukaryotes, namely multicellular forms, and represent cases where a distinct β-rich fold has been recruited for DNA binding (p53) (133) or where cell cycle regulatory elements, such as the helical cyclin box domain, have been recruited for protein–protein interactions important in the regulation of repair (Rb) (134).

Obviously, the present discussion provides only a rough sketch of the comparative aspects of the eukaryotic repair system and by no means accounts for its entire complexity, particularly with respect to the connections with transcription and the cell cycle. There is no doubt that only some of the components providing these connections have been identified to date and, furthermore, the results of our analysis point out uncertainties with regard to the actual functions of some important eukaryotic repair proteins (Table 2). For example, the product of the yeast RNC1 gene has been reported to be a DNase essential for most recombination events (135,136). However, comparative sequence analysis clearly indicates that the RNC1 protein consists of a SAM-dependent methyltransferase domain and an S1-like RNA-binding domain, suggesting an RNA methylase activity and leaving no room for a nuclease domain (Table 2; data not shown). In a similar conundrum, the yeast RAD6–RAD18 heterodimer involved in the post-replicative bypass of UV lesions has been reported to possess not only the ubiquitin-conjugating activity (intrinsic in RAD6) but also a DNA-dependent ATPase activity (137). Not only, however, does neither of the two proteins involved show any resemblance of known ATPases, but there seems to be no unaccounted for globular domain to accommodate such an activity. Further experimental studies are indispensable to solve these contradictions.

SOME GENERAL TRENDS IN THE EVOLUTION OF REPAIR SYSTEMS

The evolutionary analysis of the repair machinery reveals some general features that may reflect the selection forces behind the evolution of the repair system. The most striking aspect of the phyletic distribution of repair system is the near lack of universal components. There seem to be at least three primary evolutionary forces that shape the repair systems.

THE PRESSURES OF EXTERNAL AND INTERNAL ENVIRONMENTS

The environment and evolutionary history have profoundly affected the evolution of repair systems. Bacterial pathogens not only have small genomes, which may ease the requirement for sophisticated repair systems, but also thrive in environments where evolvability appears to be advantageous and selected for. More specifically, rapid evolution of variant antigens through replication errors and extensive recombination appears to be critical for the survival of these organisms. In these systems, the selective pressure to evade the host immune system may counterbalance the deleterious effect of 'weak', error-prone repair. As a consequence, the genomes of *Mycoplasma*, *Helicobacter*, *Borrelia* and *Treponema* lack many of the repair components present in such free-living bacteria as *Synechocystis*, *E.coli or B.subtilis* (Table 1). Even among these pathogens, however, there are considerable differences in the repertoires of the repair enzymes as demonstrated by a detailed comparison of the *Borrelia* and *Treponema* genomes (G.Subramanian, L.Aravind and E.V.Koonin, unpublished observations). Specifically, *Borrelia* that shows particularly prominent antigenic variation (138) and therefore could be expected to undergo selection for evolvability seems to have lost several genes coding for enzymes of RER that are seen in *Treponema*. This illustrates the dramatic effect of the specific lifestyle on the repair systems even among relatively close bacterial species.

Conversely, the free-living organisms, for which highly efficient repair is a must, tend to recruit additional repair enzymes. Examples of such recruits include DNA polymerase II in *E.coli* (139), DNA polymerases of the X-family in some bacteria, as well as a host of novel predicted repair enzymes in the *Mycobacteria* (116) (Table 1; Fig. 1A). Furthermore, the free-living organisms that are subject to rapid changes in the environment have an added layer of complexity in the form of the regulation of repair at the transcription level by specialized regulators, such as LexA and UmuD, that in turn are rapidly activated by damaged DNA. Free-living organisms with larger genomes seem to generate the necessary genomic variation and sustain evolvability via error-prone repair mechanisms, such as the UmuC system in bacteria (104) and apparently the analogous

system based on REV3 and REV1 in yeast, which provides error-prone translesion repair (140). A clear-cut case showing the role of the external environment in the evolution of repair enzymes is the photolyase that requires visible light and is involved primarily in the direct repair of pyrimidine photodimers (141); this enzyme is invariably missing in species that are not likely to face light, such as pathogenic bacteria and the hyperthermophilic archaea. It is particularly striking that the photosynthetic cyanobacterium *Synechocystis*, for which light exposure is evidently maximal, encodes three distinct versions of the photolyase.

The internal environment within the cell is also critical for the evolution of the repair systems as becomes clear from the nature of changes seen in eukaryotes compared to the prokaryotes. Eukaryotes have histones with basic tails complexed with the DNA and a higher order chromatin structure that is significantly more complicated than its prokaryotic counterparts (142). The evolution of these structures placed additional barriers to the repair enzymes interacting with the damaged DNA and led to the concomitant evolution of specific structural elements that provide the connection between the repair machinery and the chromatin, such as the adaptor domains discussed above. Furthermore, the tight coupling of the repair machinery with transcription (7) seen in eukaryotes appears to have co-evolved with the components of eukaryotic chromatin and cell cycle regulation. Such central components of this coupling as Rb and the cyclins that as subunits of TFIIH, participate in both repair and transcription could have evolved from TFIIB-like proteins, which also have the cyclin fold (134), and given their conservation in archaea and eukaryotes, should have been already present in their common ancestor. It is further imaginable that the cyclins originally involved in the transcription-repair coupling could have been recruited for their present role in cell cycle control, given the requirement for the recognition of damaged DNA prior to the commencement of the S-phase and the progression of cell division.

The rise of multicellularity may have mounted pressure for further developments in the coupling of repair and transcription. The need to have tissue-specific genes transcriptionally activated in the presence of damaged DNA may have provided the selective pressure for the evolution of multiple mechanisms linking the two processes. This could have been the driving force behind the evolution of such proteins as BRCA1, which participates in repair in conjunction with RAD51 (the recA ortholog) (143) and is also a part of the transcriptional machinery through its association with RNA polymerase II (144,145). While BRCA1-like proteins are seen in both plants and animals and thus seem to have an ancient origin, the transcription factor p53 is seen so far only in the coelomate animals. Three paralogs of this family are represented in mammals where there is evidence for a central role of p53 in repair (146). In addition to its function in transcription, p53 also directly associates with repair proteins, such as the recA homologs (147) and the xth-like Ap endonuclease ref-1 (148), and is involved in cell cycle arrest in response to DNA damage (149). This is a striking example of an entirely novel protein that may have evolved in only a subset of multicellular organisms, in response to the selective pressures for the coordination of transcription, repair and cell cycle.

HORIZONTAL GENE TRANSFER AND DIFFERENTIAL GENE LOSS

Another major but hitherto under-appreciated aspect of the evolution of the repair systems seems to be the role of lateral gene transfer and genomic chimerism in the generation of their diversity. As discussed above, many of the eukaryotic repair proteins clearly can be traced to bacterial and archaeal roots. Those shared with the archaea (Table 2) may come directly from the ancestor of the nuclear genome. By contrast, those repair proteins that are shared by eukaryotes and bacteria to the exclusion of the archaea, may have entered the eukaryotic lineage through horizontal transfer from the organellar (mitochondrial or chloroplast) genomes (Tables 1 and 2). Examples of this phenomenon include the RecQ family helicases, the MMR system and the BRCT domain. Routes of bacterial gene influx other than the mitochondria–nuclear transfer cannot be ruled out, particularly when very early stages of eukaryotic evolution are considered. Genomic data from other eukaryotes, particularly early branching ones, such as for example *Plasmodium*, may help in understanding the process more clearly. In each of these cases, the invasion of the eukaryotic lineage seems to have been followed by extensive duplication leading to the expansion of each of these families in eukaryotes. This must have been driven by the existence of new niches in the internal environment of the eukaryotic cell (see above), in which these proteins could acquire new, though related to the original ones, functions. A clear case of horizontal acquisition of a repair system by an archaeon from a bacterial source is the UvrABCD system in *Methanobacterium*. The RAD25/Ercc3 helicase family may represent a much less frequent case of the opposite direction of horizontal transfer. The domain conservation and phylogenetic tree analysis suggest horizontal transfer from the eukaryotes to certain bacterial species, such as *Mycobacterium leprae* and *Treponema pallidum*. The potential participation of transposable elements in the evolution of certain repair proteins, such as the xthA/AP endonucleases, is raised by their relationships with the retroelement endonucleases (150).

On many occasions, horizontal gene transfer events are difficult to distinguish from lineage-specific gene loss. In fact, this dilemma arises each time when an episodic distribution of a gene or a whole system is observed. The RecBCD exonuclease is a good example of such a situation (see above). It appears likely that the actual history of any particular repair system should have included both horizontal gene transfer and differential gene loss. The difficulties in deciphering the exact scenario notwithstanding, it is clear that the evolution of repair systems is a dramatic manifestation of the genome plasticity. Conceivably, horizontal gene transfer and lineage-specific gene loss could have been more rampant in the history of repair than in other cases, such as for example the evolution of the translation apparatus (though see 151,152), because while repair as such is essential for any organism, many of the specific repair systems can be inactivated without an immediate lethal effect (1).

PREADAPTATION: WHICH REPAIR SYSTEMS HAVE BEEN INHERITED FROM THE CENANCESTOR?

Evidently, the present layout of the repair systems in the three superkingdoms of life depends to a considerable extent on what had been inherited by each of them from their last common ancestor (the cenancestor). The comparison between bacteria, archaea and eukaryotes discussed above may help in at least partially defining this common heritage. All interpretations in this area are necessarily speculative. Nevertheless, the most parsimonious solution, considering all the data from complete genomes, is that the cenancestor at least encoded a RecA-like recombinase, a few

helicases and nucleases of the conserved superfamilies, and ABC superfamily ATPases of the SbcC/SMC2 family. This leads to a reasonably confident estimate of approximately 10 types of repair protein domains in the cenancestor. The evolution of the conserved repair pathways by vertical descent, however, appears to be largely restricted to each single superkingdom of life. This pattern is reminiscent of the profound differences in the core replicative enzymes, such as the DNA polymerases, ligases and replicative helicases and ATPases, in the archaeal/eukaryotic and bacterial lineages and is in sharp contrast with the universal conservation of the translation machinery. As discussed previously, these observations put together may suggest that the cenancestor had an RNA genome (153). If so, how does one account for the about 10 universal families of repair proteins? The general explanation is that they already had functions in an RNA-based ancestral cell—most of these conserved families of nucleases and helicases have members with RNA substrates. It is notable in this regard that the most common nucleic acid-binding module in repair proteins, HhH, is represented by both RNA-binding and DNA-binding versions. It is of further interest that the version found in eukaryotic and archaeal orthologs of RecA shows the closest similarity to the RNA-binding version in the NusA protein (see above). This raises the possibility of direct recruitment of RNA interacting proteins for roles in DNA replication and repair. This might have happened on multiple occasions in evolution—like, for example, in the Werner's syndrome protein that contains a RecQ helicase inserted into an RNase D-like domain (Fig. 1A). The XP-E and SNM1 proteins and their homologs involved in polyA processing (see above) provide additional notable examples of a connection between repair and RNA metabolism.

CONTINUING EVOLUTION OF DNA REPAIR PROTEINS

The diversity of the repair systems in different lineages indicates that they have been undergoing continuous evolution up until the terminal branches of the phylogenetic radiation. The helicases–nuclease fusions that are seen on multiple occasions in different lineages and apparently have evolved independently are a good case in point. One example is the human WRND protein, in which the helicase–nuclease fusion is not detectable in yeasts or in other animal lines, such as *C.elegans*, suggesting a relatively recent event. Similar fusions of the pol III ε subunit-like nuclease domain with the DinG helicase in *B.subtilis* (32) and with the Uri nuclease domain and the BRCT domain, respectively, in two mycobacterial proteins also are indicative of continuous generation of novel repair proteins by domain fusion.

Another notable feature observed in certain lineages is the disruption of the catalytic motifs detected on several independent occasions in ATPases and nucleases (Fig. 1A and B). In spite of the disruption of these motifs, which in all likelihood, abolishes the enzymatic activity, the domains retain detectable sequence similarity to the respective active enzymes, spanning their entire length and indicative of structural conservation. It appears likely that the original, active enzymatic domains possessed, in addition, an adaptor or regulatory function, and this is what had been preserved by selection during the subsequent evolution, after the enzymatic activity had been made obsolete by the propagation of structurally and functionally related ATPases and nucleases.

CONCLUSIONS

Comparative analysis of DNA repair systems, made possible by the availability of multiple complete genome sequences, suggests a remarkably complex picture of evolution, contingent on the external and internal environment and replete with domain shuffling, horizontal gene transfer, and lineage-specific gene loss events. Repair systems rely on a limited set of conserved domains but the number of universal repair proteins with domain architectures that are at least partially conserved across the three domains of life is very small, and there is no orthology at the level of systems and pathways. By contrast, a much greater level of conservation is observed within each of the three superkingdoms of life. The dramatic complexity of the eukaryotic repair system in terms of the number of components can be traced to the intimate connections with chromatin dynamics and cell cycle control. The repair mechanisms in archaea have not been characterized in detail. Comparative analysis readily identifies a number of candidate repair proteins but is inadequate in terms of reconstructing entire pathways. While it seems fairly safe to infer the layout of the repair systems of poorly characterized bacteria on the basis of orthologous relationships between their genes and those from well-characterized model organisms (primarily *E.coli*), understanding the archaeal systems still requires the critical body of experimental data. Similarly, a lot remains to be learnt about the details of the relationships between repair, chromatin and cell cycle in eukaryotes. It is our hope that the present analysis of the relationships between repair domains and proteins, particularly the description of previously undetected domains, will help in the rational design of experiments to further our understanding of this essential cellular function.

REFERENCES

1 Friedberg,E.C., Walker,G.C. and Siede,W. (1995) *DNA Repair and Mutagenesis*. ASM Press, Washington, DC.
2 Koonin,E.V., Mushegian,A.R. and Rudd,K.E. (1996) *Curr. Biol.*, **6**, 404–416.
3 Hanawalt,P.C. (1995) *Mutat. Res.*, **336**, 101–113.
4 Sancar,A. (1996) *Annu. Rev. Biochem.*, **65**, 43–81.
5 Modrich,P. and Lahue,R. (1996) *Annu. Rev. Biochem.*, **65**, 101–133.
6 Wood,R.D. (1996) *Annu. Rev. Biochem.*, **65**, 135–167.
7 Friedberg,E.C. (1996) *Annu. Rev. Biochem.*, **65**, 15–42.
8 Minnick,D.T. and Kunkel,T.A. (1996) *Cancer Surv.*, **28**, 3–20.
9 Wood,R. (1997) *J. Biol. Chem.*, **272**, 23465–23468.
10 Parikh,S.S., Mol,C.D. and Tainer,J.A. (1997) *Structure*, **5**, 1543–1550.
11 Walker,D.R. and Koonin,E.V. (1997) *ISMB*, **5**, 333–339.
12 Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.
13 Wootton,J.C. and Federhen,S. (1996) *Methods Enzymol.*, **266**, 554–571.
14 Tatusov,R.L., Altschul,S.F. and Koonin,E.V. (1994) *Proc. Natl Acad. Sci. USA*, **91**, 12091–12095.
15 Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) *Nucleic Acids Res.*, **22**, 4673–4680.
16 Schuler,G.D., Altschul,S.F. and Lipman,D.J. (1991) *Proteins*, **9**, 180–190.
17 Neuwald,A.F., Liu,J.S. and Lawrence,C.E. (1995) *Protein Sci.*, **4**, 1618–1632.
18 Rost,B. and Sander,C. (1994) *Proteins*, **19**, 55–72.
19 Rost,B., Schneider,R. and Sander,C. (1997) *J. Mol. Biol.*, **270**, 471–480.
20 Fitch,W.M. (1970) *System. Zool.*, **19**, 99–106.
21 Fitch,W.M. (1995) *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, **349**, 93–102.
22 Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) *Science*, **278**, 631–637.
23 Tatusov,R.L., Mushegian,A.R., Bork,P., Brown,N.P., Hayes,W.S., Borodovsky,M., Rudd,K.E. and Koonin,E.V. (1996) *Curr. Biol.*, **6**, 279–291.
24 Henikoff,S., Greene,E.A., Pietrokovski,S., Bork,P., Attwood,T.K. and Hood,L. (1997) *Science*, **278**, 609–614.
25 Doolittle,R.F. (1995) *Annu. Rev. Biochem.*, **64**, 287–314.
26 Doolittle,R.F. and Bork,P. (1993) *Sci. Am.*, **269**, 50–56.

27 Harris,P.V., Mazina,O.M., Leonhardt,E.A., Case,R.B., Boyd,J.B. and Burtis,K.C. (1996) *Mol. Cell Biol.*, **16**, 5764–5771.
28 Gorbalenya,A.E. and Koonin,E.V. (1990) *J. Mol. Biol.*, **213**, 583–591.
29 Gorbalenya,A.E. and Koonin,E.V. (1993) *Curr. Opin. Struct. Biol.*, **3**, 419–429.
30 Braithwaite,D.K. and Ito,J. (1993) *Nucleic Acids Res.*, **21**, 787–802.
31 Doolittle,R.F., Johnson,M.S., Husain,I., Van Houten,B., Thomas,D.C. and Sancar,A. (1986) *Nature*, **323**, 451–453.
32 Moser,M.J., Holley,W.R., Chatterjee,A. and Mian,I.S. (1997) *Nucleic Acids Res.*, **25**, 5110–5118.
33 Lieber,M.R. (1997) *Bioessays*, **19**, 233–240.
34 Aravind,L. and Koonin,E.V. (1998) *Nucleic Acids Res.*, **26**, 3746–3752.
35 Ramotar,D. (1997) *Biochem. Cell Biol.*, **75**, 327–336.
36 Carrell,H.L., Glusker,J.P., Burger,V., Manfre,F., Tritsch,D. and Biellmann,J.F. (1989) *Proc. Natl Acad. Sci. USA*, **86**, 4440–4444.
37 Lavie,A., Allen,K.N., Petsko,G.A. and Ringe,D. (1994) *Biochemistry*, **33**, 5469–5480.
38 Holm,L. and Sander,C. (1997) *Proteins*, **28**, 72–82.
39 Yajima,H., Takao,M., Yasuhira,S., Zhao,J.H., Ishii,C., Inoue,H. and Yasui,A. (1995) *EMBO J.*, **14**, 2393–2399.
40 Takao,M., Yonemasu,R., Yamamoto,K. and Yasui,A. (1996) *Nucleic Acids Res.*, **24**, 1267–1271.
41 Lin,J.J., Phillips,A.M., Hearst,J.E. and Sancar,A. (1992) *J. Biol. Chem.*, **267**, 17693–17700.
42 Lin,J.J. and Sancar,A. (1992) *J. Biol. Chem.*, **267**, 17688–17692.
43 Derbyshire,V., Kowalski,J.C., Dansereau,J.T., Hauer,C.R. and Belfort,M. (1997) *J. Mol. Biol.*, **265**, 494–506.
44 Sijbers,A.M., van der Spek,P.J., Odijk,H., van den Berg,J., van Duin,M., Westerveld,A., Jaspers,N.G., Bootsma,D. and Hoeijmakers,J.H. (1996) *Nucleic Acids Res.*, **24**, 3370–3380.
45 Doherty,A.J., Serpell,L.C. and Ponting,C.P. (1996) *Nucleic Acids Res.*, **24**, 2488–2497.
46 Tomkinson,A.E., Bardwell,A.J., Tappe,N., Ramos,W. and Friedberg,E.C. (1994) *Biochemistry*, **33**, 5305–5311.
47 Yu,M., Souaya,J. and Julin,D.A. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 981–986.
48 Haijema,B.J., Venema,G. and Kooistra,J. (1996) *J. Bacteriol.*, **178**, 5086–5091.
49 Champoux,J.J. (1998) *Prog. Nucleic Acids Res. Mol. Biol.*, **60**, 111–132.
50 Evans,B.R., Chen,J.W., Parsons,R.L., Bauer,T.K., Teplow,D.B. and Jayaram,M. (1990) *J. Biol. Chem.*, **265**, 18504–18510.
51 Tirumalai,R.S., Healey,E. and Landy,A. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 6104–6109.
52 Ilyina,T.V. and Koonin,E.V. (1992) *Nucleic Acids Res.*, **20**, 3279–3285.
53 Thayer,M.M., Ahern,H., Xing,D., Cunningham,R.P. and Tainer,J.A. (1995) *EMBO J.*, **14**, 4108–4120.
54 Morozov,V., Mushegian,A.R., Koonin,E.V. and Bork,P. (1997) *Trends Biochem. Sci.*, **22**, 417–418.
55 Moolenaar,G.F., Franken,K.L., Dijkstra,D.M., Thomas-Oates,J.E., Visse,R., van de Putte,P. and Goosen,N. (1995) *J. Biol. Chem.*, **270**, 30508–30515.
56 Callebaut,I. and Mornon,J.P. (1997) *FEBS Lett.*, **400**, 25–30.
57 Bork,P., Hofmann,K., Bucher,P., Neuwald,A.F., Altschul,S.F. and Koonin,E.V. (1997) *FASEB J.*, **11**, 68–76.
58 Masson,M., Niedergang,C., Schreiber,V., Muller,S., Menissier-de Murcia,J. and de Murcia,G. (1998) *Mol. Cell Biol.*, **18**, 3563–3571.
59 Tomkinson,A.E. and Mackey,Z.B. (1998) *Mutat Res.*, **407**, 1–9.
60 Yu,X., Wu,L.C., Bowcock,A.M., Aronheim,A. and Baer,R. (1998) *J. Biol. Chem.*, **273**, 25388–25392.
61 Hofmann,K. and Bucher,P. (1995) *Trends Biochem. Sci.*, **20**, 347–349.
62 Ohta,K., Nicolas,A., Furuse,M., Nabetani,A., Ogawa,H. and Shibata,T. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 646–651.
63 Sun,Z., Hsiao,J., Fay,D.S. and Stern,D.F. (1998) *Science*, **281**, 272–274.
64 Aravind,L. and Koonin,E.V. (1998) *Trends Biochem. Sci.*, **23**, 284–286.
65 Jonsson,Z.O. and Hubscher,U. (1997) *Bioessays*, **19**, 967–975.
66 Jonsson,Z.O., Hindges,R. and Hubscher,U. (1998) *EMBO J.*, **17**, 2412–2425.
67 Krishna,T.S., Kong,X.P., Gary,S., Burgers,P.M. and Kuriyan,J. (1994) *Cell*, **79**, 1233–1243.
68 Thelen,M.P., Onel,K. and Holloman,W.K. (1994) *J. Biol. Chem.*, **269**, 747–754.
69 Parker,A.E., Van de Weyer,I., Laus,M.C., Oostveen,I., Yon,J., Verhasselt,P. and Luyten,W. (1998) *J. Biol. Chem.*, **273**, 18332–18339.
70 Kong,X.P., Onrust,R., O'Donnell,M. and Kuriyan,J. (1992) *Cell*, **69**, 425–437.
71 Griffiths,D.J., Barbet,N.C., McCready,S., Lehmann,A.R. and Carr,A.M. (1995) *EMBO J.*, **14**, 5812–5823.
72 Oku,T., Ikeda,S., Sasaki,H., Fukuda,K., Morioka,H., Ohtsuka,E., Yoshikawa,H. and Tsurimoto,T. (1998) *Genes Cells*, **3**, 357–369.
73 Roca,A.I. and Cox,M.M. (1997) *Prog. Nucleic Acids Res. Mol. Biol.*, **56**, 129–223.
74 Bianco,P.R., Tracy,R.B. and Kowalczykowski,S.C. (1998) *Front. Biosci.*, **3**, d570–d603.
75 Sargentini,N.J. and Smith,K.C. (1986) *Radiat. Res.*, **107**, 58–72.
76 Song,Y. and Sargentini,N.J. (1996) *J. Bacteriol.*, **178**, 5045–5048.
77 Koonin,E.V., Tatusov,R.L. and Rudd,K.E. (1996) In Neidhardt,F.C.,III, R.C., Ingraham,J.L., Lin,E.C.C., Low,K.B., Magasanik,B., Resnikoff,W.S., Riley,M., Schaechter,M. and Umbarger,H.E. (eds), *Escherichia coli and Salmonella typhimurium*. ASM Press, Washington, DC, Vol. 2, pp. 2203–2217.
78 Harrington,J.J. and Lieber,M.R. (1994) *Genes Dev.*, **8**, 1344–1355.
79 Harrington,J.J. and Lieber,M.R. (1994) *EMBO J.*, **13**, 1235–1246.
80 Shen,B., Qiu,J., Hosfield,D. and Tainer,J.A. (1998) *Trends Biochem. Sci.*, **23**, 171–173.
81 Sharma,R.C. and Smith,K.C. (1987) *J. Bacteriol.*, **169**, 4559–4564.
82 Peterson,C.L. (1994) *Cell*, **79**, 389–392.
83 Michaelis,C., Ciosk,R. and Nasmyth,K. (1997) *Cell*, **91**, 35–45.
84 Koonin,E.V. (1994) *Protein Sci.*, **3**, 356–358.
85 Sharples,G.J. and Leach,D.R. (1995) *Mol. Microbiol.*, **17**, 1215–1217.
86 Connelly,J.C., Kirkham,L.A. and Leach,D.R.F. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 7969–7974.
87 West,S.C. (1997) *Annu. Rev. Genet.*, **31**, 213–244.
88 van Gool,A.J., Shah,R., Mezard,C. and West,S.C. (1998) *EMBO J.*, **17**, 1838–1845.
89 Eisen,J.A., Kaiser,D. and Myers,R.M. (1997) *Nat. Med.*, **3**, 1076–1078.
90 Bergerat,A., de Massy,B., Gadelle,D., Varoutas,P.C., Nicolas,A. and Forterre,P. (1997) *Nature*, **386**, 414–417.
91 Mushegian,A.R., Bassett,D.E.,Jr, Boguski,M.S., Bork,P. and Koonin,E.V. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 5831–5836.
92 Reenan,R.A. and Kolodner,R.D. (1992) *Genetics*, **132**, 975–985.
93 Hanada,K., Ukita,T., Kohno,Y., Saito,K., Kato,J. and Ikeda,H. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 3860–3865.
94 Davey,S., Han,C.S., Ramer,S.A., Klassen,J.C., Jacobson,A., Eisenberger,A., Hopkins,K.M., Lieberman,H.B. and Freyer,G.A. (1998) *Mol. Cell Biol.*, **18**, 2721–2728.
95 Ellis,N.A., Groden,J., Ye,T.Z., Straughen,J., Lennon,D.J., Ciocci,S., Proytcheva,M. and German,J. (1995) *Cell*, **83**, 655–666.
96 Lombard,D.B. and Guarente,L. (1996) *Trends Genet.*, **12**, 283–286.
97 Aravind,L. and Koonin,E.V. (1998) *Trends Biochem. Sci.*, **23**, 17–19.
98 Dianov,G. and Lindahl,T. (1994) *Curr. Biol.*, **4**, 1069–1076.
99 Dianov,G., Sedgwick,B., Daly,G., Olsson,M., Lovett,S. and Lindahl,T. (1994) *Nucleic Acids Res.*, **22**, 993–998.
100 Clark,A.J. (1991) *Biochimie*, **73**, 523–532.
101 Webb,B.L., Cox,M.M. and Inman,R.B. (1997) *Cell*, **91**, 347–356.
102 Aravind,L., Leipe,D.D. and Koonin,E.V. (1998) *Nucleic Acids Res.*, **26**, 4205–4213.
103 Little,J.W. (1993) *J. Bacteriol.*, **175**, 4943–4950.
104 Smith,B.T. and Walker,G.C. (1998) *Genetics*, **148**, 1599–1610.
105 Selby,C.P. and Sancar,A. (1993) *Science*, **260**, 53–58.
106 Selby,C.P. and Sancar,A. (1995) *J. Biol. Chem.*, **270**, 4882–4889.
107 Palmer,B.R. and Marinus,M.G. (1994) *Gene*, **143**, 1–12.
108 Ban,C. and Yang,W. (1998) *EMBO J.*, **17**, 1526–1534.
109 Friedberg,E.C. (1991) *Mol. Microbiol.*, **5**, 2303–2310.
110 Lehmann,A.R. (1998) *Bioessays*, **20**, 146–155.
111 Sung,P., Bailly,V., Weber,C., Thompson,L.H., Prakash,L. and Prakash,S. (1993) *Nature*, **365**, 852–855.
112 Weeda,G., van Ham,R.C., Vermeulen,W., Bootsma,D., van der Eb,A.J. and Hoeijmakers,J.H. (1990) *Cell*, **62**, 777–791.
113 Troelstra,C., van Gool,A., de Wit,J., Vermeulen,W., Bootsma,D. and Hoeijmakers,J.H. (1992) *Cell*, **71**, 939–953.
114 Koonin,E.V. (1993) *Nucleic Acids Res.*, **21**, 1497.
115 Gerring,S.L., Spencer,F. and Hieter,P. (1990) *EMBO J.*, **9**, 4347–4358.
116 Poterszman,A., Lamour,V., Egly,J.M., Moras,D., Thierry,J.C. and Poch,O. (1997) *Trends Biochem. Sci.*, **22**, 418–419.
117 Eisen,J.A., Sweder,K.S. and Hanawalt,P.C. (1995) *Nucleic Acids Res.*, **23**, 2715–2723.
118 Bork,P. and Koonin,E.V. (1993) *Nucleic Acids Res.*, **21**, 751–752.
119 Muzzin,O., Campbell,E.A., Xia,L., Severinova,E., Darst,S.A. and Severinov,K. (1998) *J. Biol. Chem.*, **273**, 15157–15161.

120 Johnson,R.E., Henderson,S.T., Petes,T.D., Prakash,S., Bankmann,M. and Prakash,L. (1992) *Mol. Cell Biol.*, **12**, 3807–3818.
121 Bang,D.D., Verhage,R., Goosen,N., Brouwer,J. and van de Putte,P. (1992) *Nucleic Acids Res.*, **20**, 3925–3931.
122 Habraken,Y., Sung,P., Prakash,L. and Prakash,S. (1994) *J. Biol. Chem.*, **269**, 31342–31345.
123 Rapic Otrin,V., Kuraoka,I., Nardo,T., McLenigan,M., Eker,A.P., Stefanini,M., Levine,A.S. and Wood,R.D. (1998) *Mol. Cell Biol.*, **18**, 3182–3190.
124 Wolter,R., Siede,W. and Brendel,M. (1996) *Mol. Gen. Genet.*, **250**, 162–168.
125 Aravind,L. (1998) *Silico Biol.*, **1**, 8.
126 Borden,K.L. and Freemont,P.S. (1996) *Curr. Opin. Struct. Biol.*, **6**, 395–401.
127 Henning,K.A., Li,L., Iyer,N., McDaniel,L.D., Reagan,M.S., Legerski,R., Schultz,R.A., Stefanini,M., Lehmann,A.R., Mayne,L.V. *et al. (*1995) *Cell*, **82**, 555–564.
128 Masutani,C., Sugasawa,K., Yanagisawa,J., Sonoyama,T., Ui,M., Enomoto,T., Takio,K., Tanaka,K., van der Spek,P.J., Bootsma,D. *et al. (*1994) *EMBO J.*, **13**, 1831–1843.
129 Allen,J.B., Zhou,Z., Siede,W., Friedberg,E.C. and Elledge,S.J. (1994) *Genes Dev.*, **8**, 2401–2415.
130 Zhou,Z. and Elledge,S.J. (1993) *Cell*, **75**, 1119–1127.
131 Carr,A.M. (1997) *Curr. Opin. Genet. Dev.*, **7**, 93–98.
132 Sherr,C.J. (1996) *Science*, **274**, 1672–1677.
133 Jeffrey,P.D., Gorina,S. and Pavletich,N.P. (1995) *Science*, **267**, 1498–1502.
134 Gibson,T.J., Thompson,J.D., Blocker,A. and Kouzarides,T. (1994) *Nucleic Acids Res.*, **22**, 946–952.
135 Chow,T.Y., Perkins,E.L. and Resnick,M.A. (1992) *Nucleic Acids Res.*, **20**, 5215–5221.
136 Moore,P.D., Simon,J.R., Wallace,L.J. and Chow,T.Y. (1993) *Curr. Genet*, **23**, 1–8.
137 Bailly,V., Lauder,S., Prakash,S. and Prakash,L. (1997) *J. Biol. Chem.*, **272**, 23360–23365.
138 Koomey,M. (1997) *Curr. Biol.*, **7**, R538–R540.
139 Bonner,C.A., Hays,S., McEntee,K. and Goodman,M.F. (1990) *Proc. Natl Acad. Sci. USA*, **87**, 7663–7667.
140 Lawrence,C.W. and Hinkle,D.C. (1996) *Cancer Surv.*, **28**, 21–31.
141 Hearst,J.E. (1995) *Science*, **268**, 1858–1859.
142 Lamond,A.I. and Earnshaw,W.C. (1998) *Science*, **280**, 547–553.
143 Scully,R., Chen,J., Plug,A., Xiao,Y., Weaver,D., Feunteun,J., Ashley,T. and Livingston,D.M. (1997) *Cell*, **88**, 265–275.
144 Anderson,S.F., Schlegel,B.P., Nakajima,T., Wolpin,E.S. and Parvin,J.D. (1998) *Nature Genet.*, **19**, 254–256.
145 Chapman,M.S. and Verma,I.M. (1996) *Nature*, **382**, 678–679.
146 Osada,M., Ohba,M., Kawahara,C., Ishioka,C., Kanamaru,R., Katoh,I., Ikawa,Y., Nimura,Y., Nakagawara,A., Obinata,M. and Ikawa,S. (1998) *Nat. Med.*, **4**, 839–843.
147 Buchhop,S., Gibson,M.K., Wang,X.W., Wagner,P., Sturzbecher,H.W. and Harris,C.C. (1997) *Nucleic Acids Res.*, **25**, 3868–3874.
148 Jayaraman,L., Murthy,K.G., Zhu,C., Curran,T., Xanthoudakis,S. and Prives,C. (1997) *Genes Dev.*, **11**, 558–570.
149 Ko,L.J. and Prives,C. (1996) *Genes Dev.*, **10**, 1054–1072.
150 Feng,Q., Moran,J.V., Kazazian,H.H.,Jr and Boeke,J.D. (1996) *Cell*, **87**, 905–916.
151 Ibba,M., Bono,J.L., Rosa,P.A. and Soll,D. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 14383–14388.
152 Koonin,E.V. and Aravind,L. (1998) *Curr. Biol.*, **8**, R266–R269.
153 Mushegian,A.R. and Koonin,E.V. (1996) *Proc. Natl Acad. Sci. USA*, **93**, 10268–10273.

NOTE ADDED IN PROOF

While this manuscript was being processed for publication, several interesting findings on proteins and domains involved in repair and evolution of repair systems have been published. The crystal structures of two important domains involved in repair have been determined, namely the N-terminal domain of *E.coli* MutL protein [Ban,C. and Yang,W. (1998) *Cell*, **95**, 541–552] and the BRCT domain from the human XRCC1 protein (Zhang,X., Morera,S., Bates,P.A., Whitehead,P.C., Coffer,A.I., Hainbucher,K., Nash,R.A., Sternberg,M.J., Lindahl,T. and Freemont,P.S. (1998) *EMBO J.*, **17**, 6404–6411]. The MutL structure was found to be highly similar to those of the ATPase domain of DNA gyrase and HSP90, which confirms the earlier predictions; the ATPase activity of MutL has been demonstrated experimentally. The BRCT domain was found to possess a new fold. One of the RecA family ATPases from *Synechocystis* sp., which contains a duplication of the ATPase domain, has been shown to participate in the generation of circadian oscillation in this cyanobacterium [Ishiura,M., Kutsuna,S., Aoki,S., Iwasaki,H., Andersson,C.R., Tanabe,A., Golden,S.S., Johnson,C.H. and Kondo,T. (1998) *Science*, **281**, 1519–1523]. This is in agreement with the notion of likely non-repair functions of some ATPases of the RecA family and suggests that the highly conserved archaeal orthologs of this cyanobacterial protein also might be involved in signal transduction rather than in repair. A detailed phylogenetic analysis of the MutS protein family was published; the results support a very early duplication of MutS, with subsequent functional diversification of the duplicates [Eisen,J.A. (1998) *Nucleic Acids Res.*, **26**, 4291–4300]. Analysis of the genome of the hyperthermophilic bacterium *A.aeolicus* revealed a number of likely horizontal transfers from archaea; these include a considerable set of proteins implicated in repair, such as a RecA-type ATPase, two distinct DNA ligases and several nucleases [Aravind,L., Tatusov,R.L., Wolf,Y.I., Walker,D.R. and Koonin,E.V. (1998) *Trends Genet.*, **14**, 442–444].