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. 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.

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...
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 phylogenetic 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 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 Biotechnology 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 phylectic 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 phylectic 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, nuclease 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 RecA 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<10^{-3}, 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 RX3[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 D339 and D466 in E.coli UvrC (Fig. 2C) abolished the 5′ incision but did not affect 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

<table>
<thead>
<tr>
<th>Protein</th>
<th>Function / Activity</th>
<th>Pathwaya</th>
<th>Phytophagous distributionb</th>
<th>Bacteriaa</th>
<th>Escherichia</th>
<th>Eukaryac</th>
<th>Domainc</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>PhcB</td>
<td>Photolysin</td>
<td>DR</td>
<td>+ + + + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Flavins and 8-hydroxy-2-deazafuran-dependent light receptor domains</td>
<td>Markedly episodic distribution; among Gs., found in B. subtilis and Streptomyces but not in L. rhamnosus or Mobocrella, some clones in these bacteria (see text).</td>
</tr>
<tr>
<td>Ada</td>
<td>O-6-alkylguanine, O-4-alkylguanine DNA alkylation, remove alkyl groups of many types, transcription activator</td>
<td>DR</td>
<td>+ + + + + + + + + + + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>CIC2, Zs argin-Ascaris family HTH-like methyltransferase</td>
<td>Arthral and mammalian homologues have only the methyltransferase domain; panelling of Ogt</td>
</tr>
<tr>
<td>Ogt</td>
<td>O-6-methylguanine DNA methyltransferase</td>
<td>DR</td>
<td>+ + + + + + + + + + + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>methyltransferase</td>
<td>Panelling of Ada without the additional domains</td>
</tr>
<tr>
<td>MutT</td>
<td>8-oxo-GTPase</td>
<td>DR</td>
<td>+ + + + + + + + + + + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>MutT (Nudix) hydrolase</td>
<td>A vast family of pyrophosphohydrolases; some of the orthologous relationships should be considered provisional</td>
</tr>
<tr>
<td>DnaD</td>
<td>dCTPase</td>
<td>DR</td>
<td>+ + + + + + + + + + + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>dCTPase domain, domains related to dCTP-domain (E. coli) (151)</td>
<td>Among the pyrophosphohydrolases, found in E. coli, and in B. subtilis, also recorded by several bacterial genomes, crenobacteria, and bacteriophages. Paper showing of DnaA. Universal in Archaea but rare in bacteria</td>
</tr>
<tr>
<td>dCTPase</td>
<td>dCTP-deaminase</td>
<td>DR</td>
<td>+ + + + + + + + + + + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>dCTPase domain; domains related to dCTP-domain (E. coli) (151)</td>
<td>Family of bacterial glucosyltransferases belonging to endonuclease III (Nhe, Mmr). The bacterial ortholog has an N-terminal fusion of an Ada-like CIC2-like finger</td>
</tr>
<tr>
<td>ARA</td>
<td>3-methylaminol, 3-methylguanine, O-2-methylcytosine, O-2-methylthymine DNA glycosylase II</td>
<td>DR, BER</td>
<td>+ + + + + + + + + + + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Glycosylase+BHH</td>
<td>Glycosylase+BHH</td>
</tr>
<tr>
<td>AAB</td>
<td>Unknown</td>
<td>DR</td>
<td>+ + + + + + + + + + + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Novel predicted hydrolase domain</td>
<td>New family found only in E. coli, and Caulobacter, and in a divergent form in Mycobacterium among bacteria, but also in animals and plants (not in yeast) and in the polygenic of plant DNA strains of the cosexual and telomerase strains (E. amarum and E. v. koonin, unpublished observations)</td>
</tr>
<tr>
<td>MutY</td>
<td>8-oxoguanine DNA glycosylase &amp; AP lyase, A-G mismatch DNA glycosylase</td>
<td>BER</td>
<td>+ + + + + + + + + + + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Glycosylase/endonuclease+BHH; cytosine-rich motif</td>
<td>Family of a CA motif in the cell, and in Mycobacteria, and in a divergent form in Mycobacterium among bacteria, but also in animals and plants (not in yeast) and in the polygenic of plant DNA strains of the cosexual and telomerase strains (E. amarum and E. v. koonin, unpublished observations)</td>
</tr>
<tr>
<td>Nih</td>
<td>Endonuclease III &amp; thymine glycol DNA glycosylase</td>
<td>BER</td>
<td>+ + + + + + + + + + + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Distinct glycosylase/endonuclease+BHH/C4 little finger motif</td>
<td>Novel</td>
</tr>
<tr>
<td>MutM</td>
<td>Formamidopyrimidine &amp; 8-oxoguanine DNA glycosylase</td>
<td>BER</td>
<td>+ + + + + + + + + + + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Distinct glycosylase/endonuclease+BHH/C4 little finger motif</td>
<td>Novel</td>
</tr>
<tr>
<td>Fpg</td>
<td>Endonuclease VIII</td>
<td>BER</td>
<td>+ + + + + + + + + + + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Distinct glycosylase/endonuclease+BHH/C4 little finger motif</td>
<td>Distinct glycosylase/endonuclease+BHH/C4 little finger motif</td>
</tr>
<tr>
<td>Nss</td>
<td>Endonuclease IV</td>
<td>BER</td>
<td>+ + + + + + + + + + + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Distinct glycosylase/endonuclease+BHH/C4 little finger motif</td>
<td>Distinct glycosylase/endonuclease+BHH/C4 little finger motif</td>
</tr>
<tr>
<td>CyoP</td>
<td>DNA polymerase I</td>
<td>BER</td>
<td>+ + + + + + + + + + + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>RNA polymerase-type 5'-3' exonuclease+3'-5' exonuclease + polymerase</td>
<td>The RNA polymerases have a N-terminal SF1 helicase domain (see text and Fig. 1A)</td>
</tr>
<tr>
<td>Tag</td>
<td>3-methyladenine DNA glycosylase I</td>
<td>BER</td>
<td>+ + + + + + + + + + + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Distinct glycosylase domain</td>
<td>Distinct glycosylase domain</td>
</tr>
<tr>
<td>Ung</td>
<td>uracil DNA glycosylase</td>
<td>BER</td>
<td>+ + + + + + + + + + + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Distinct glycosylase domain</td>
<td>Distinct glycosylase domain</td>
</tr>
<tr>
<td>XbaI</td>
<td>Endonuclease XbaI</td>
<td>BER</td>
<td>+ + + + + + + + + + + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Distinct glycosylase domain</td>
<td>Distinct glycosylase domain</td>
</tr>
<tr>
<td>RadC</td>
<td>Predicted DNA-binding protein</td>
<td>BER</td>
<td>+ + + + + + + + + + + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>BHH+uncharacterized conserved domain</td>
<td>BHH+uncharacterized conserved domain</td>
</tr>
<tr>
<td>RadA/</td>
<td>Predicted ATP-dependent protease</td>
<td>NER</td>
<td>+ + + + + + + + + + + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>C2a finger-like domain + RecA family ATPase+ Lectin-like protein</td>
<td>The protease appears to be active in some forms of the protest and reactive in others. Stand-alone forms of the protease domain are found in other proteins.</td>
</tr>
<tr>
<td>Smc</td>
<td></td>
<td>BER</td>
<td>+ + + + + + + + + + + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>C2a finger-like domain + RecA family ATPase+ Lectin-like protein</td>
<td>C2a finger-like domain + RecA family ATPase+ Lectin-like protein</td>
</tr>
<tr>
<td>Mll</td>
<td>transcription repair coupling factor; helicase</td>
<td>NER</td>
<td>+ + + + + + + + + + + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>SFIF helicase (degraded) + SFIF helicase</td>
<td>Fusion with a degraded wobbling helicase in the N terminal. Universal in bacteria except Mycoplasma</td>
</tr>
<tr>
<td>UvaA</td>
<td>ATPase, DNA binding</td>
<td>NER</td>
<td>+ + + + + + + + + + + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>ABC-family ATPase/Finger (see Fig. 1B)</td>
<td>Universal in bacteria; among the archaea, found only in Methanobrevibacter and Methanococcus. Possibly interacting with UvaC with the common UVR family DNA.</td>
</tr>
<tr>
<td>UvrB</td>
<td>Helicase</td>
<td>NER</td>
<td>+ + + + + + + + + + + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>SFIF helicase + UVR domain</td>
<td>UVR helicase</td>
</tr>
<tr>
<td>UvrC</td>
<td>Nuclease II; Invites unwinding from a nick</td>
<td>NER</td>
<td>+ + + + + + + + + + + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>5'-endonuclease+3'-exonuclease+BHH+UVR helicase</td>
<td>UVR helicase</td>
</tr>
<tr>
<td>UvrD</td>
<td>Nuclease I; Invites unwinding from a nick</td>
<td>NER</td>
<td>+ + + + + + + + + + + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>SFIF helicase</td>
<td>Universal in bacteria</td>
</tr>
<tr>
<td>MsdL</td>
<td>predicted ATPase</td>
<td>mMM, mMM</td>
<td>+ + + + + + + + + + + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>HSP90 family ATPase</td>
<td>ATPase of the HSP90 family group</td>
</tr>
<tr>
<td>MsdS</td>
<td>ATPase</td>
<td>mMM, mMM</td>
<td>+ + + + + + + + + + + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>ABC-superfamily ATPase</td>
<td>ATPase of the HSP90 family group</td>
</tr>
<tr>
<td>MsdH</td>
<td>Endonuclease</td>
<td>mMM, mMM</td>
<td>+ + + + + + + + + + + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Sau3A-like restriction endonuclease domain</td>
<td>Sau3A-like restriction endonuclease domain</td>
</tr>
<tr>
<td>Var</td>
<td>O-acetylguanine DNA glycosylase</td>
<td>mMM, mMM</td>
<td>+ + + + + + + + + + + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Adenine-specific DNA methylase</td>
<td>Adenine-specific DNA methylase</td>
</tr>
<tr>
<td>XerA/Xe</td>
<td>Exonuclease V, large subunit</td>
<td>MM</td>
<td>+ + + + + + + + + + + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>A distinct nuclease domain</td>
<td>A distinct nuclease domain</td>
</tr>
<tr>
<td>XerB</td>
<td>Exonuclease V, small subunit</td>
<td>MM</td>
<td>+ + + + + + + + + + + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Uncorrelated domain</td>
<td>Uncorrelated domain</td>
</tr>
<tr>
<td>ShBC</td>
<td>Exonuclease III</td>
<td>mMM, mMM</td>
<td>+ + + + + + + + + + + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>3'-5' Exonuclease Sbf</td>
<td>A highly divergent version of the domain so far detected only in E. coli and I. intestinalis</td>
</tr>
<tr>
<td>Dem</td>
<td>site-specific C-C pyrimidine endonuclease</td>
<td>mMM, mMM</td>
<td>+ + + + + + + + + + + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>SAM-dependent methyltransferase</td>
<td>SAM-dependent methyltransferase</td>
</tr>
<tr>
<td>DnrP</td>
<td>Specific function unknown (predicted nucleolytic/transaminase)</td>
<td>mMM, mMM</td>
<td>+ + + + + + + + + + + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Nucleobase transfer+DHH</td>
<td>Panelling of UvaA. Among archaea, so far only in Sulfolobus</td>
</tr>
<tr>
<td>SbcC</td>
<td>Exonuclease subunit, predicted ATPase</td>
<td>REK</td>
<td>+ + + + + + + + + + + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>ABC family ATPase with coiled coils</td>
<td>Nearly universal but missing in Mycoplasma. Helicase, 1 helix, in spite of the preponderance of the coiled-coil structure, orthologs could be shown through distinct signature motifs</td>
</tr>
<tr>
<td>SbcD</td>
<td>Exonuclease</td>
<td>REK</td>
<td>+ + + + + + + + + + + +</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Calcium-dependent phosphohydrolase domain</td>
<td>Calcium-dependent phosphohydrolase domain</td>
</tr>
</tbody>
</table>
Table 1. Continued

| RecA | recordase, ssDNA dependent ATPase, activator of LexA autoreplication |
| RecB | Helicase/mosaicase |
| RecC | Helicase/mosaicase |
| RecD | Helicase/mosaicase |
| RecF | predicted ATPase, required for resumption of DNA replication at displaced replication forks |
| RecG | Holliday junction-specific DNA helicase, branch migration mediator |
| RecJ | Nucleate |
| RecN | predicted ATPase |
| RecO | “anti-sf” domain; stabilization of RecA filaments; ATP-independent, Rec-A-like strand association activity |
| RecQ | helicase; suppressor of illegitimate recombination |
| RecR | required for resumption of DNA replication at displaced replication forks |
| RecA endonuclease/Holliday junction |
| RecB | Holliday junction resolver; ATPase subunit of a helicase |
| RecC | Holliday junction resolver; endonuclease |
| RecE | exonuclease |
| RecT | nuclease protein |
| DnaG | predicted helicase; SOS indoor |
| LexA | transcriptional regulator, antiproteinase |
| PolB | DNA polymerase II |
| UmuC | in conjunction with umuD and recA, facilitates transition DNA synthesis |
| UmuD | in conjunction with umuC and recA, facilitates transition DNA synthesis |
| DnaE | polymerase subunit of the DNA polymerase III holoenzyme |
| DnaQ | 3’-5’ exonuclease subunit of the DNA polymerase III holoenzyme |
| DnaJ | DNA ligase |
| SsrA | Single-stranded binding protein |

RADI/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 *Saccharomyces 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 ERKX2SD 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 (Fig. 1A). The remarkable feature of this protein family is that in archaea,
### Table 2. Yeast DNA repair systems: conservation in completely sequenced genomes

<table>
<thead>
<tr>
<th>Gene</th>
<th>Function / Activity</th>
<th>Pathway</th>
<th>Organisms*</th>
<th>Arc</th>
<th>Bac</th>
<th>Domains</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAD1</td>
<td>Single-strand DNA endonuclease; Cuts at duplex 5'-3' single-strand junctions</td>
<td>NER, RER</td>
<td>+ + + (+)</td>
<td>Disrupted SFH helicase + nuclease (ERCC4 family) + HhH</td>
<td>The yeast protein lacks the C-terminal HhH domains found in the orthologs from other species including E. pombe</td>
<td>See text and Figs. 1A and 2E</td>
<td></td>
</tr>
<tr>
<td>RAD10</td>
<td>Single-strand DNA endonuclease subunit (RAD1-RAD10 complex)</td>
<td>NER, RER</td>
<td>+ + - -</td>
<td>Disrupted nucleases (ERCC4 family) + HhH</td>
<td>Some of the eukaryotic members contain large non globular inserts. In bacteria, the orthologous domain is fused to DNA polymerase I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAD2</td>
<td>Single-strand DNA endonuclease 5'-3' exonuclease; cleaves at duplex 5'-3' single-strand junctions</td>
<td>NER</td>
<td>+ + + (+)</td>
<td>5'-3' exonuclease + HhH</td>
<td>XP-C ortholog. Yeast also has a paralogous gene which is conserved in E. pombe</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAD3</td>
<td>5'-3' DNA and DNA-RNA helicase; Pol II basal transcription factor</td>
<td>NER</td>
<td>+ + (+)</td>
<td>SFH helicase</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAD4</td>
<td>Possibly involved in repair; transcription coupling and in the repair of cycle connection; exact role unknown</td>
<td>NER</td>
<td>- + - -</td>
<td>No known domains detectable</td>
<td>Leucine rich repeat protein.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAD7</td>
<td>Involved in NER as a complex with RAD16</td>
<td>NER</td>
<td>(+) - - -</td>
<td>Leucine-rich repeats</td>
<td>Leucine rich repeat protein.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAD16</td>
<td>DNA helicase</td>
<td>NER</td>
<td>+ + (+)</td>
<td>All eukaryotic orthologs have a RING finger</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAD14</td>
<td>Damage specific DNA-binding protein</td>
<td>NER</td>
<td>+ + - -</td>
<td>CIC2 Zn finger + HDC2 finger-like motif</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAD23</td>
<td>Provides connection between NER and ubiquitin-dependent proteasome pathways</td>
<td>NER</td>
<td>+ + - -</td>
<td>Ubiquitin-lysine characterized domain</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAD25 (SSL2, UV512)</td>
<td>5'-3' DNA helicase; Pol II basal transcription protein</td>
<td>NER</td>
<td>+ + (+)</td>
<td>Uncharacterized conserved domain + SFH helicase</td>
<td>Superfamily II helicase with a specific conserved N-terminal domain</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSL1</td>
<td>TP1H 44 KD subunit</td>
<td>NER</td>
<td>+ + - -</td>
<td>von Willebrand factor A domain + Zn finger.</td>
<td>Homolog of the protosomal subunit SS152, and L. Aravind, unpublished observations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPB1</td>
<td>TP1H 63 KD subunit</td>
<td>NER</td>
<td>+ + - -</td>
<td>Novel repetitive motif</td>
<td>Archael and bacterial ATP-dependent ligase lack the BRCT domain</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDC9</td>
<td>DNA Ligase</td>
<td>NER</td>
<td>+ + (+)</td>
<td>ATP-dependent ligase + BRCT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MMS19</td>
<td>Transcriptional complex apparently active through interaction with TP1H</td>
<td>NER</td>
<td>+ - - -</td>
<td>Leucine-rich repeats</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNM1 (PSO2)</td>
<td>Protein required for DNA cross-link repair; predicted nuclease</td>
<td>NER</td>
<td>+ + (+)</td>
<td>Metallo β-lactamase</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAD6-epistasis group</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAD5 (REV2)</td>
<td>Helicase</td>
<td>TLR*</td>
<td>+ + (+)</td>
<td>SFH/SWI helicases + RING finger</td>
<td>Domain architecture analogous to</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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, archaenal, 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>
<thead>
<tr>
<th>RAD52 epistasis group</th>
<th>Connection between mRNA processing and repair</th>
<th>RAD6</th>
<th>Ubiquitin-conjugating enzyme, corrects repair with protein degradation, forms DNA-binding heterodimers with RAD18</th>
<th>RER, + + + + Ubiquitin-conjugating enzyme</th>
<th>RAD16 and other eukaryotic SNF1</th>
<th>TLR, and novel pathways</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAD9</td>
<td>DNA damage checkpoint component</td>
<td>RER</td>
<td>- - - - 2 C-terminal BRCT domains</td>
<td>So far no orthologs detectable</td>
<td>Genuine orthologs seen only in other fungi.</td>
<td></td>
</tr>
<tr>
<td>RAD21/DSS1</td>
<td>DNA helicase involved in the RAD1-dependent NER pathway</td>
<td>RER</td>
<td>(+) (+) (+) (+) (+) (+) SF1 helicase.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>REV1</td>
<td>DNA polymerase ζ subunit, predicted template-independent nucleotidyl transferase</td>
<td>TLR</td>
<td>+ (+) (+) (+) Nucleotidyl transferase + HhB+BRCT</td>
<td>Prokaryotic homologs, such as DnaP, lack the BRCT domain but have HhB motifs.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>REV3</td>
<td>DNA polymerase ζ elongation subunit</td>
<td>TLR</td>
<td>(+) + (+) Superfamily 8 DNA polymerase + unique Cys-rich domain</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>REV7</td>
<td>DNA polymerase ζ subunit</td>
<td>TLR</td>
<td>(+) (+) Homolog domain</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>REV10</td>
<td>Novel rad18 dependent pathway component</td>
<td>Novel pathway</td>
<td>(+) (+) (+) Nucleotidyl transferase</td>
<td>A distinct paralog of REV1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PS045PRP(9)</td>
<td>Connection between mRNA processing and repair</td>
<td>RER</td>
<td>+ + + + WD40 repeats</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 2. Continued

<table>
<thead>
<tr>
<th>RAD50</th>
<th>Chromatin modifying ATPase</th>
<th>RER</th>
<th>+ + + + ABC superfamily ATPase, inserted-coil coiled domains</th>
<th>See text and Fig. 1B</th>
<th>Paralog of RAD50</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAD51</td>
<td>ATPase</td>
<td>RER</td>
<td>+ + + + Modified HhB-related ATPase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAD52</td>
<td>Forms a complex involved in strand exchange with RAD55 and RAD57</td>
<td>RER</td>
<td>- - - - Uncharacterized conserved domain</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAD53/SKP1</td>
<td>Protein Ser/Thr kinase</td>
<td>RER</td>
<td>+ + + + FHA+Ser/Thr kinase+FHA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAD54</td>
<td>Helicase involved in strand exchange in conjunction with RAD51</td>
<td>RER</td>
<td>+ (+) (+) (+) SNF/SWI type SF1 helicase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAD55</td>
<td>ATPase</td>
<td>RER</td>
<td>- + + + Modified HhB-related ATPase</td>
<td>See text and Fig. 1B</td>
<td></td>
</tr>
<tr>
<td>RAD57</td>
<td>ATPase</td>
<td>RER</td>
<td>- + + + Modified HhB-related ATPase</td>
<td>See text and Fig. 1B</td>
<td></td>
</tr>
<tr>
<td>RAD24</td>
<td>ATPase; DNA damage checkpoint component interacting with RAD17 and MEC3</td>
<td>RER</td>
<td>(+) (+) (+) (+) RF-C type AAA superfamily ATPase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRE11</td>
<td>3'-5' exoribonuclease and endonuclease, as a complex with RAD50, involved non-homologous joining of DNA ends</td>
<td>RER</td>
<td>+ + + + Nuclease of the calcineurin-like fold</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAD9</td>
<td>Involved in double-strand break repair, function unknown</td>
<td>RER</td>
<td>- (+) - Uncharacterized conserved domain</td>
<td>Paralog of RAD2</td>
<td></td>
</tr>
<tr>
<td>XR52</td>
<td>Involved in double-strand break repair, as a complex with RAD50 and MRE11</td>
<td>RER</td>
<td>- - - - Divergent FHA domain + coiled-coil.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAD17</td>
<td>DNAase; DNA damage checkpoint component interacting with RAD24 and MEC3</td>
<td>RER</td>
<td>(+) (+) (+) (+) PCNA fold domain</td>
<td>See text and Fig. 2F</td>
<td></td>
</tr>
<tr>
<td>MEC3</td>
<td>DNA damage checkpoint component interacting with RAD17 and RAD24</td>
<td>- - - - No identifiable structural features</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MMS21</td>
<td>Function unknown</td>
<td>?</td>
<td>+ - - - Zn-coordinating Cys-His cluster</td>
<td>No detectable homologs</td>
<td>No detectable homologs</td>
</tr>
<tr>
<td>REC114</td>
<td>Function unknown</td>
<td>RER</td>
<td>- - - - Coiled-coil</td>
<td></td>
<td></td>
</tr>
<tr>
<td>REC102</td>
<td>Function unknown</td>
<td>RER</td>
<td>- - - - No identifiable structural features</td>
<td>No detectable homologs</td>
<td>Ortholog in S. pombe</td>
</tr>
<tr>
<td>REC103/SK8</td>
<td>Function unknown; predicted adapter</td>
<td>RER</td>
<td>- - - - WD40 repeats</td>
<td></td>
<td></td>
</tr>
<tr>
<td>REC104</td>
<td>Function unknown</td>
<td>RER</td>
<td>- - - - No identifiable structural features</td>
<td>No detectable homologs</td>
<td>Ortholog in S. pombe</td>
</tr>
<tr>
<td>RNC1</td>
<td>Claimed to be a nuclease (133)</td>
<td>RER</td>
<td>- + + + Divergent S1 domain + SAM-dependent methyltransferase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPO41</td>
<td>Double-strand break introducing endonuclease</td>
<td>RER</td>
<td>+ + + + Divergent Toprim domain (100)</td>
<td>The archael orthologs are subunits of Toporionase VL</td>
<td></td>
</tr>
<tr>
<td>MEC1</td>
<td>DNA-dependent protein kinase; DNA damage checkpoint component; phosphotyrosine DCC1</td>
<td>RER</td>
<td>+ + + + Lipid kinase superfamily domain</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DDC-1</td>
<td>Checkpoint sensor</td>
<td>RER</td>
<td>- - - - Distantly related to Rad-9 from S. pombe</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Other repair proteins

<table>
<thead>
<tr>
<th>OGG1</th>
<th>8-oxoguanine DNA glycosylase</th>
<th>DR, BER</th>
<th>- + + + glycosylase domain+HhB</th>
<th>Family of a helical glycosylases homologous to endonuclease III</th>
<th>Family of a helical glycosylases homologous to endonuclease III; close paralogs of NG1 and distant paralog of OGG1</th>
</tr>
</thead>
<tbody>
<tr>
<td>NG1</td>
<td>8-oxoguanine DNA glycosylase</td>
<td>DR, BER</td>
<td>+ + + + glycosylase domain+HhB</td>
<td>Family of a helical glycosylases homologous to endonuclease III</td>
<td>Family of a helical glycosylases homologous to endonuclease III; close paralogs of NG1 and distant paralog of OGG1</td>
</tr>
<tr>
<td>NG2</td>
<td>8-oxoguanine DNA glycosylase</td>
<td>DR, BER</td>
<td>+ + + + glycosylase domain+HhB</td>
<td>Family of a helical glycosylases homologous to endonuclease III</td>
<td>Family of a helical glycosylases homologous to endonuclease III; close paralogs of NG1 and distant paralog of OGG1</td>
</tr>
<tr>
<td>PFI1</td>
<td>5'-3' helicase involved in mitochondrial repair</td>
<td>?</td>
<td>(+) (+) (+) SFI helicase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAD26</td>
<td>Helicase involved in transcription-repair coupling</td>
<td>NER</td>
<td>+ (+) (+) SNF/SWI type SFI helicase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KEM1/BRAS5</td>
<td>5'-3' Nuclease</td>
<td>RER</td>
<td>+ + + + Novel nucleosome</td>
<td>In bacteria, the orthologous domain is fused to DNA polymerase I</td>
<td>In bacteria, the orthologous domain is fused to DNA polymerase I</td>
</tr>
<tr>
<td>RAD27</td>
<td>Single-strand DNA endonuclease</td>
<td>NER, MMR7</td>
<td>+ + + + 5'-3' exoribonuclease+HhB</td>
<td>In bacteria, the orthologous domain is fused to DNA polymerase I</td>
<td>In bacteria, the orthologous domain is fused to DNA polymerase I</td>
</tr>
<tr>
<td>DSK-7</td>
<td>5'-3' Nuclease</td>
<td>MMR</td>
<td>+ + + + 5'-3' exoribonuclease+HhB</td>
<td>In bacteria, the orthologous domain is fused to DNA polymerase I</td>
<td>In bacteria, the orthologous domain is fused to DNA polymerase I</td>
</tr>
<tr>
<td>EOX-1</td>
<td>5'-3' Nuclease</td>
<td>MMR</td>
<td>+ + + + 5'-3' exoribonuclease+HhB</td>
<td>In bacteria, the orthologous domain is fused to DNA polymerase I</td>
<td>In bacteria, the orthologous domain is fused to DNA polymerase I</td>
</tr>
<tr>
<td>PFR1</td>
<td>Photolysase</td>
<td>DR</td>
<td>- + + + Flavin and 8-hydroxy-5-deazafavin-dependent light receptor domain</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Ce, C. elegans; Hs, H. sapiens; Arc, archaea; Bac, bacteria. See also footnotes to Table 1. **TLR, trans-lesion repair.
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 computational methods (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).
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, Paramaecium bursaria Chlorella virus; Aa, A.aloicus; Aae, Alcaligenes eutrophus; At, A.fulgidus; Aaem, Allomyces macrogynus; At, Arabidopsis thaliana; Bb, Borrelia burgdorferi; Bs, B.subtilis; Ce, C.elegans; Cel, Chlorogonium elongatum; Cegl, 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; Ngo, Neisseria gonorrhoeae; Pa, Podospora anserina; Pf, Photorhabdus; Ph, Phorichosia; Pv, Phaseolus vulgaris; Rsph, Rhodopseudomonas spheroides; Sag, Streptococcus agalactiae; Sc, S.cerevisiae; Sp, S.pombe; Ss, Synechocystis sp.; St, Streptococcus thermophilus; Tp, Tpallidum; 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 and cell cycle regulation and may provide the critical connection 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 in both 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.
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).

**PHYLETIC DISTRIBUTION AND EVOLUTION OF REPAIR SYSTEMS**

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 of 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 *srs* 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 genome 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 Uni 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 RuvABC(C) complex, which is the Holliday junction resolvase and the key component of bacterial BER (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 (BNM domain in Fig. 1A), which is also found in eukaryotic chromatin-associated proteins, such as BS69 and BR140 (L.Arvind, 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 nucleus, 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.
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 Smr 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 archaean 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 nuclease. 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 phylogenetic distribution—in addition to its conservation in eukaryotes, it is also present in the archaean *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 nucleosome 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 phylogenetic 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 SW1 family helicases and other proteins like RAD18, the WD40 domains, such as the RING finger (126) in some of the SWI 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 phylogenetic 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.

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 Synchocystis, 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 their 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 TFIH, 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 UvrABC 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, HH1, 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 helicase–nuclease fusions that are seen on multiple occasions in different 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, HH1, 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.

**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**


**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., Morena, S., Bates, P.A., Whitehead, P.C., Coffier, A., Hainbucher, K., Nash, R.A., Stemberg, 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. and 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 archaean 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., Tatussov, R.L., Wolf, Y.I., Walker, D.R. and Koonin, E.V. (1998) Trends Genet., 14, 442–444].