© 1998 Oxford University Press Nucleic Acids Research, 1998, Vol. 26, No. 18 4291–4300

A phylogenomic study of the MutS family of proteins

Jonathan A. Eisen*

Department of Biological Sciences, Stanford University, Stanford, CA 94305-5020, USA

Received January 23, 1998; Revised May 28, 1998; Accepted July 28, 1998

ABSTRACT

The MutS protein of *Escherichia coli* plays a key role in the recognition and repair of errors made during the replication of DNA. Homologs of MutS have been found in many species including eukaryotes, Archaea and other bacteria, and together these proteins have been grouped into the MutS family. Although many of these proteins have similar activities to the *E.coli* MutS, there is significant diversity of function among the MutS family members. This diversity is even seen within species; many species encode multiple MutS homologs with distinct functions. To better characterize the MutS protein family, I have used a combination of phylogenetic reconstructions and analysis of complete genome sequences. This phylogenomic analysis is used to infer the evolutionary relationships among the MutS family members and to divide the family into subfamilies of orthologs. Analysis of the distribution of these orthologs in particular species and examination of the relationships within and between subfamilies is used to identify likely evolutionary events (e.g. gene duplications, lateral transfer and gene loss) in the history of the MutS family. In particular, evidence is presented that a gene duplication early in the evolution of life resulted in two main MutS lineages, one including proteins known to function in mismatch repair and the other including proteins known to function in chromosome segregation and crossing-over. The inferred evolutionary history of the MutS family is used to make predictions about some of the uncharacterized genes and species included in the analysis. For example, since function is generally conserved within subfamilies and lineages, it is proposed that the function of uncharacterized proteins can be predicted by their position in the MutS family tree. The uses of phylogenomic approaches to the study of genes and genomes are discussed.

INTRODUCTION

The ability to recognize and repair mismatches in DNA after replication has occurred has been well documented in many species. While some such mismatch repair (MMR) is carried out by pathways that repair only specific DNA replication errors, most is performed by broad specificity ‘general’ MMR pathways. The most extensively studied general MMR system is the MutHLS pathway of the bacterium *Escherichia coli* (see 1,2 for review). In the first critical step in this pathway, the MutS protein (in the form of a dimer) binds to the site of a mismatch in double-stranded DNA. Through a complex interaction between MutS, MutL and MutH, a section of the newly replicated DNA strand (and thus the strand with the replication error) at the location of the mismatch bound by MutS is targeted for removal. Other proteins complete the repair process: the section of DNA that has been targeted is removed and degraded, a patch is synthesized using the complementary strand as a template and the patch is ligated into place resulting in a section of double-stranded DNA without mismatches.

The ability of the MutHLS pathway to repair many types of replication errors is due to the broad specificity of MutS recognition and binding. Since MutS binds to many types of base:base mismatches, the MutHLS pathway can repair many types of base misincorporation errors. Similarly, since MutS binds to heteroduplex loops (in which one strand contains extra-helical bases) the MutHLS pathway can repair frameshift replication errors. This ability to repair loops was somewhat surprising since this pathway was originally characterized as being involved in repairing mismatches. The repair of loops is particularly important in the regulation of the stability of microsatellites (loci that contain 1–10 bp tandem repeats). Microsatellites are prone to a special class of frameshift replication errors due to a process known as slip-strand mispairing (SSM). This process leads to the generation of loops of one or more copies of repeat unit (3,4). The MutHLS pathway helps keep microsatellite mutation rates in check by repairing many of the loops generated by SSM (5). While the specificity of MutS binding (and thus the MutHLS pathway) is quite broad, it is not uniform. For example, MutS does not bind C:C mismatches well and therefore the misincorporation of a C opposite a C will not be repaired well by the MutHLS pathway (6). Binding of MutS to heteroduplex loops is also not uniform. MutS only binds loops of up to four bases in size and only binds well to those up to three bases in size (7). Thus frameshift errors are only repaired if they produce loops of four bases or smaller. Since loops generated by SSM in microsatellites are usually one repeat unit in size, microsatellites with repeats >4 bp are highly unstable in *E.coli*. The non-uniformity of MutS recognition causes the MutHLS pathway to influence not only the mutation rate, but also the mutation spectrum.

The overall scheme of the MutHLS pathway (mismatch recognition, strand discrimination and excision and resynthesis) is conserved in the general MMR systems of other species (1).

*Tel: +1 650 723 2425; Fax: +1 650 725 1848; Email: jeisen@leland.stanford.edu

However, the degree of conservation of specific details varies greatly between the different steps in the process. Some steps (e.g. strand recognition) do not even use the same general mechanism between species. Others (e.g. exonucleolytic degradation) are similar in biochemical mechanism but make use of non-homologous proteins in different species. Nevertheless, some of the specific details of the MMR process are highly conserved. In particular, homologs of MutL and MutS are required for general MMR in all species examined and these proteins function in much the same way as the *E.coli* MutL and MutS (1). The conservation of MutS between species makes the specificity of MMR similar to that of *E.coli*. As with the *E.coli* MutHLS pathway, all characterized general MMR systems can repair both mismatches and loops. Incidentally, this is what led to the discovery that hereditary non-polyposis colon cancer (HNPCC) can be caused by defects in MMR (8). Cells from patients with HNPCC showed exceptionally high levels of microsatellite instability, due to defects in loop repair.

While the ability to repair both loops and mismatches is conserved, the specificity of other species MMR is not identical to that of *E.coli*. As with *E.coli*, dissecting the specificity of MMR in other species requires dissection of the binding preferences of MutS (or in these cases MutS homologs). However, in many cases the comparison to the *E.coli* MutS is complicated. For example, the best-studied eukaryotic MMR system is that of the yeast *Saccharomyces cerevisiae*. Unlike *E.coli*, *S.cerevisiae* encodes six MutS homologs, referred to as MutS Homolog (MSH) proteins (9). The best characterized of these are MSH2, MSH3 and MSH6 which are involved in MMR in the nucleus. These proteins are combined to create two distinct heterodimers; one for recognizing and repairing base:base mismatches and loops of one to two bases (composed of MSH2 and MSH6), and one for recognizing and repairing larger loops (composed of MSH2 and MSH3) (4,10). Thus, since MSH2 is in both heterodimers, it is required for all MMR in the nucleus, while MSH3 and MSH6 provide the specificity for the type of replication error recognized. The roles of the other MutS homologs in *S.cerevisiae* are not as well understood. MSH1 is involved in the repair of mismatches in mitochondrial DNA, although its exact function is not known (11–13). MSH4 and MSH5 do not even function in MMR, but instead are involved in meiotic crossing-over and chromosome segregation (14–16). The role of MutS homologs in processes other than correction of replication errors is not surprising since mismatches can arise in a variety of cellular circumstances. The proteins in the *E.coli* MutHLS pathway also have alternative cellular roles including the regulation of interspecies recombination and the repair of certain types of DNA damage (1,17). It may be that some of the multiple roles of the *E.coli* MutS have been divided up among the many *S.cerevisiae* MutS homologs.

Mismatch recognition and repair in humans and other animals is quite similar to that of *S.cerevisiae* (18,19; A.Villanuve, personal communication). Preliminary studies suggest that this is also true for plants (20). These similarities suggest that the complex MMR system of *S.cerevisiae* was established prior to the divergence of animal, fungal and plant ancestors. While studies of MMR in model species like humans, *S.cerevisiae* and *E.coli* are likely to continue, most new information about the MutS family of proteins is coming in the form of sequence data. Sequences of MutS homologs continue to pour into sequence databases, most without any accompanying functional information. An important new source of these sequences has been genome projects and the results coming out of these projects are somewhat surprising. For example, two MutS homologs have been found in many bacterial species as a result of bacterial genome projects (21,22), but it is not known if their functions are distinct. In addition, some bacteria do not encode any MutS homologs and some species do not encode any MutS homologs, while others encode a MutS homolog but no MutL homolog (23).

How can one make sense out of the ever-expanding MutS family, the diversity of MutS proteins within particular species, and these unusual distribution patterns in complete genome sequences? In this paper, I describe a new type of analysis, which I refer to as phylogenomics, focused specifically on the MutS family of proteins. This analysis provides insight into the evolution of the MutS protein family and the diversity of functions within and between species. In addition, it allows improved predictions of the functions of uncharacterized genes in the MutS family, and the likely phenotypes of species for which complete genomes are available. Such a phylogenomic analysis can be useful to studies of any gene family.

MATERIALS AND METHODS

The sequences of previously characterized MutS-like proteins were downloaded from the National Center for Biotechnology Information (NCBI) databases (accession numbers are given in Table 1). Additional members of the MutS family were searched for using the blast (24), blast2 and PSI-blast (25) computer programs. Databases searched included the NCBI non-redundant database and unpublished, nearly complete genome sequences of *Deinococcus radiodurans* and *Treponema pallidum* from The Institute for Genomic Research (personal communication) and *Streptococcus pyogenes* and *Neisseria gonorrhoeae* from University of Oklahoma (B.A.Roe, S.Clifton and D.W.Dyer, personal communication).

Protein sequences were aligned using the *clustalw* (26) and *clustalx* (27) multiple sequence alignment programs with some manual adjustment using the GDE computer software package (28,29). Regions of ambiguity in this alignment were determined by comparison to alternative alignments generated using modifications of the alignment parameters (such as different gap penalties).

Phylogenetic trees were generated from the sequence alignments using the PAUP* program (30) on a PowerBook 3400/180. Parsimony analysis was conducted using the *heuristic search* algorithm. The total branch length of trees was quantified using either an identity matrix, a PAM250 matrix or a MutS-specific matrix (based on the frequency of particular amino acid substitutions in the evolution of the MutS protein family as estimated by the MacClade program; 31). Multiple runs searching for the shortest tree were conducted for each matrix. Distance-based phylogenetic trees were generated by the neighbor-joining (32) and UPGMA algorithms using estimated evolutionary calculated from the matrices described above. Bootstrap resampling was conducted by the method of Felsenstein (33). Character state analysis for the study of gene loss was conducted using the MacClade computer program (31).

RESULTS AND DISCUSSION

The publication in 1995 of the complete genome sequence of the bacterium *Haemophilus influenzae* (34) signaled the beginning of a new era in biological research. Genome sequences provide a wealth of information not only about a single organism but also about all of the genes that they encode. As genome and other sequence data continue to pour into databases at an amazing pace, we need to develop new methods to sort out this information. In developing such methods it is important to recognize that analysis of genomes can benefit from studies of individual gene families and analysis of genome sequences can provide a great deal of information about gene families. For example, many genomes encode dozens or even hundreds of members of some multigene families. Making accurate predictions of the phenotype of these species from the genome sequence requires making accurate predictions of the functions of genes in multigene families. Similarly, a simple analysis of the presence and absence of particular genes in a genome can reveal a great deal about different multigene families. Most methods currently being used to analyze gene and genome data rely on the identification and quantification of similarity between the gene or genome of interest and those of other species. While such methods are useful, they tend to ignore the fact that biological similarities have a historical component (i.e. evolution). It is well documented that the incorporation of an evolutionary perspective can greatly benefit any comparative biological study. The benefits of the evolutionary perspective come from focusing not just on similarities and differences, but on how and why such similarities and differences arose. Therefore, I believe that studies of genes and genomes can also benefit greatly from an evolutionary focus. I refer to the combined evolutionary study of genes and genomes as phylogenomics (35,36).

I report here a phylogenomic analysis that is focused on the MutS family of proteins. The MutS family is an ideal case study for phylogenomic analysis for a variety of reasons. First, there is a good deal of functional diversity within this gene family. Thus, classifying uncharacterized genes may help improve functional predictions. In addition, this diversity of functions may have major effects on species phenotypes, in particular any phenotype related to mutation rate and pattern. Thus identifying which genes are present in a particular genome may help improve predictions of that species phenotype. Finally, as mentioned in the Introduction, there are many unusual patterns of distribution of MutS homologs in currently available complete genome sequences. I have divided the phylogenomic analysis of the MutS family into multiple sections. In the first few sections, the evolutionary history of the MutS family is inferred by analysis of genes and genomes currently available. In the remaining sections, this evolutionary information is used to place some of the studies of the members of this gene family into a useful context and also to make predictions for uncharacterized genes and species.

Identification and alignment of MutS homologs

Multiple sequence searching algorithms were used to identify proteins with extensive amino acid sequence similarity to the previously characterized members of the MutS family. To increase the likelihood of identifying all available MutS homologs, highly divergent members of the MutS family and a MutS consensus sequence were used as query sequences. In addition, the PSI-blast program was used to identify any proteins with similar motifs to other MutS-like proteins. Proteins were considered to be members of the MutS family if they showed significant sequence similarity to any of the previously identified MutS proteins, and if this similarity extended throughout the protein. All identified complete or nearly complete MutS family members are listed in Table 1.

The sequences of the proteins listed in Table 1 were aligned to each other using the *clustalw* multiple sequence alignment algorithm. This alignment was enhanced both manually and with the *clustalx* program, which allows local *clustalw* alignments to be performed within a larger alignment. (This complete alignment is available at http://www-leland.stanford.edu/~jeisen/MutS/MutS.html). The alignment reveals that there are motifs that are highly conserved among all MutS-like proteins. Most of these conserved motifs are confined to one section that is on average ~260 amino acids in length. This section can be considered the core MutS-family domain. For most of the members of the MutS family, the MutS-family domain is near the C-terminal end of each protein. The alignment of this domain is shown for a representative sample of the proteins in the MutS family in Figure 1. The levels of identity and similarity among the MutS family members ranges from 32% similarity and 18% identity between some distantly related members to 70% similarity and 60% identity between putative orthologs from human and mouse (a matrix with pairwise similarities and identities is available at the MutS website as described above). The level of similarity among all these proteins is much higher than one would expect to occur by convergence, suggesting that all these proteins share a common ancestor and thus should be considered homologs. Although all family members have a MutS-family domain, some sequence patterns were conserved only among subsets of the MutS-like proteins. These motifs may be responsible for providing specific functions to the individual MutS proteins (see below).

Phylogenetic trees of the MutS homologs

Phylogenetic trees of the proteins in the MutS family were determined from the alignment using distance and parsimony methods, each with multiple parameters (see Materials and Methods). Since each alignment position is assumed to include residues that share a common ancestry among species, regions of ambiguous alignment were excluded from the phylogenetic analysis. Regions of particularly low sequence conservation were also excluded. In total, 313 amino acid alignment positions were used (available at the MutS web site). The trees generated with the different methods and parameters were very similar in topology to each other. Therefore only one tree (the neighbor-joining tree) is shown here (Fig. 2a). Bootstrap analysis revealed that most of the patterns shown in the tree are highly robust (bootstrap values >70%). Bootstrap values of particular branches are discussed in more detail below and are shown in some of the subsequent tables and figures. Overall, the similarity of the trees generated by multiple methods and the high bootstrap values for most branches indicate that most of the patterns shown in Figure 2 are highly robust.

In addition to assessing the internal consistency of the results, it is also useful to compare the results presented here to those of other studies. Unfortunately, many previous studies of the evolution of the MutS family of proteins have not described the methods used to generate the trees and thus are not comparable to this study (e.g. 18). In addition, some studies have used multiple sequence alignment programs like *clustalw* and *pileup* to generate trees directly and thus cannot be considered reliable phylogenetic studies (37,38). There have been only two studies of the evolution of MutS homologs using standard phylogenetic methods (20,39). These studies should be considered limited because they did not include many of the more divergent members of the MutS family. Nevertheless, most of the results of these studies are similar to those reported here. Some specific differences and similarities are discussed below.

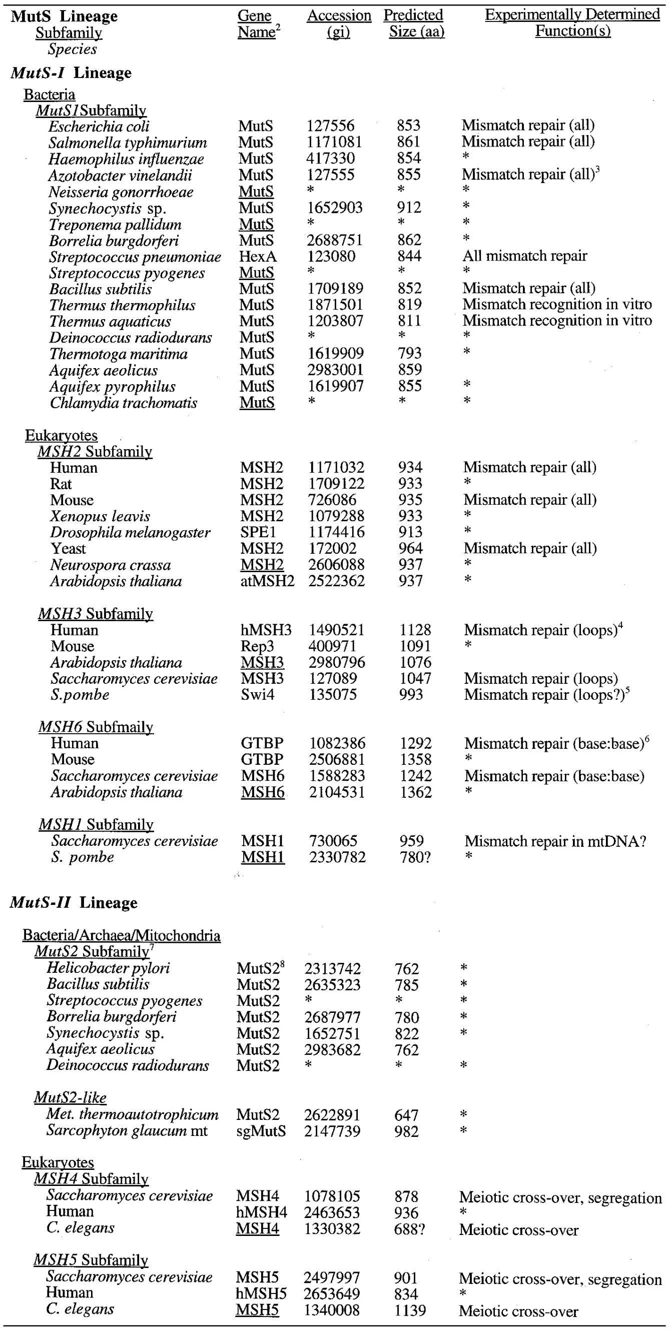
Beyond gene trees: identifying evolutionary events in the MutS family's history

As with any gene family, the phylogenetic tree of the MutS proteins simply shows the relationships among homologs. It is almost always useful to go beyond this gene tree to identify specific evolutionary events in a gene family's history. For example, identification of the types of homology (orthology, paralogy and xenology) in this tree allows the detection of the particular evolutionary event (speciation, gene duplication and lateral gene transfer, respectively) that led to the divergence of homologs. To identify these and other evolutionary events, it is necessary to integrate the gene tree with other information, such as gene function, species phenotype or species phylogeny.

Subfamilies of orthologs. As the first step in going beyond the MutS gene tree, I divided the MutS family into subfamilies that I propose represent distinct groups of orthologs (i.e. sets of genes that diverged from each other due to speciation events). Each subfamily has been given a name based on the name of one of the better-studied proteins in that group (italics are used to distinguish the subfamilies from individual proteins). The proposed subfamilies are highlighted in Figure 2B–D and the proteins in each subfamily are listed in Table 1. Some characteristics of each subfamily are given in Table 2. The assertion that these subfamilies are distinct evolutionary groups is supported by five lines of evidence: (i) each was found in trees generated by all the phylogenetic methods used; (ii) each has reasonably high bootstrap values with different methods (Table 2); (iii) the branches leading up to the subfamilies are relatively long indicating that each is evolutionarily distinct from other subfamilies; (iv) protein size is somewhat conserved within subfamilies (see Table 1); and (v) there are sequence motifs conserved within but not between subfamilies (not shown). The assertion that these evolutionarily distinct subfamilies are distinct orthologous groups is supported by two factors: (i) the phylogenetic relationships of proteins within each group are roughly congruent to the likely relationships of the species from which they come; and (ii) function has been conserved within subfamilies.

Overall, eight orthologous subfamilies were identified; six that include only proteins from eukaryotes (corresponding to the six yeast MutS homologs) and two that include only proteins from bacteria. Most of these subfamilies correspond well to groups that have been suggested previously. For example, the animal and yeast proteins in each eukaryotic subfamily have been identified as likely orthologs of each other by standard sequence similarity searches and other non-phylogenetic methods. The phylogenetic analysis simply confirms that these are indeed orthologs. The identification of two distinct bacterial subfamilies represents a novel finding [although it was suggested by Eisen *et al.* (35)]. This finding shows one of the benefits of phylogenetic analysis over standard sequence-similarity searches. In addition to the subfamilies, two proteins (one from *Methanobacterium thermoautotrophicum* and one from the mitochondrial genome of *Sarcophyton glaucum*) are closely related to the *MutS2* subfamily but they were not placed into this subfamily. Although these two genes group with the *MutS2* subfamily in every tree, it is possible that they may have been involved in lateral transfer

Table 1. Proteins in the MutS family[1]



MutS Lineage Subfamily *Species*	Gene Name[2]	Accession (gi)	Predicted Size (aa)	Experimentally Determined Function(s)
***MutS-I* Lineage**				
Bacteria				
*MutS1*Subfamily				
Escherichia coli	MutS	127556	853	Mismatch repair (all)
Salmonella typhimurium	MutS	1171081	861	Mismatch repair (all)
Haemophilus influenzae	MutS	417330	854	*
Azotobacter vinelandii	MutS	127555	855	Mismatch repair (all)[3]
Neisseria gonorrhoeae	MutS	*	*	*
Synechocystis sp.	MutS	1652903	912	*
Treponema pallidum	MutS	*	*	*
Borrelia burgdorferi	MutS	2688751	862	*
Streptococcus pneumoniae	HexA	123080	844	All mismatch repair
Streptococcus pyogenes	MutS	*	*	*
Bacillus subtilis	MutS	1709189	852	Mismatch repair (all)
Thermus thermophilus	MutS	1871501	819	Mismatch recognition in vitro
Thermus aquaticus	MutS	1203807	811	Mismatch recognition in vitro
Deinococcus radiodurans	MutS	*	*	*
Thermotoga maritima	MutS	1619909	793	*
Aquifex aeolicus	MutS	2983001	859	
Aquifex pyrophilus	MutS	1619907	855	*
Chlamydia trachomatis	MutS	*	*	*
Eukaryotes				
MSH2 Subfamily				
Human	MSH2	1171032	934	Mismatch repair (all)
Rat	MSH2	1709122	933	*
Mouse	MSH2	726086	935	Mismatch repair (all)
Xenopus leavis	MSH2	1079288	933	*
Drosophila melanogaster	SPE1	1174416	913	*
Yeast	MSH2	172002	964	Mismatch repair (all)
Neurospora crassa	MSH2	2606088	937	*
Arabidopsis thaliana	atMSH2	2522362	937	*
MSH3 Subfamily				
Human	hMSH3	1490521	1128	Mismatch repair (loops)[4]
Mouse	Rep3	400971	1091	*
Arabidopsis thaliana	MSH3	2980796	1076	
Saccharomyces cerevisiae	MSH3	127089	1047	Mismatch repair (loops)
S.pombe	Swi4	135075	993	Mismatch repair (loops?)[5]
MSH6 Subfmaily				
Human	GTBP	1082386	1292	Mismatch repair (base:base)[6]
Mouse	GTBP	2506881	1358	*
Saccharomyces cerevisiae	MSH6	1588283	1242	Mismatch repair (base:base)
Arabidopsis thaliana	MSH6	2104531	1362	*
MSH1 Subfamily				
Saccharomyces cerevisiae	MSH1	730065	959	Mismatch repair in mtDNA?
S. pombe	MSH1	2330782	780?	*
***MutS-II* Lineage**				
Bacteria/Archaea/Mitochondria				
*MutS*2 Subfamily[7]				
Helicobacter pylori	MutS2[8]	2313742	762	*
Bacillus subtilis	MutS2	2635323	785	*
Streptococcus pyogenes	MutS2	*	*	*
Borrelia burgdorferi	MutS2	2687977	780	*
Synechocystis sp.	MutS2	1652751	822	*
Aquifex aeolicus	MutS2	2983682	762	
Deinococcus radiodurans	MutS2	*	*	*
MutS2-like				
Met. thermoautotrophicum	MutS2	2622891	647	*
Sarcophyton glaucum mt	sgMutS	2147739	982	*
Eukaryotes				
MSH4 Subfamily				
Saccharomyces cerevisiae	MSH4	1078105	878	Meiotic cross-over, segregation
Human	hMSH4	2463653	936	*
C. elegans	MSH4	1330382	688?	Meiotic cross-over
MSH5 Subfamily				
Saccharomyces cerevisiae	MSH5	2497997	901	Meiotic cross-over, segregation
Human	hMSH5	2653649	834	*
C. elegans	MSH5	1340008	1139	Meiotic cross-over

[1] Only complete or nearly complete proteins are included. Additional information about each protein can be found in GenBank and at http://www-leland.stanford.edu/~jeisen/MutS/MutS.html

[2] Unnamed open reading frames are given a proposed name which is underlined.

[3] Determined by increased mutation rate in lines with defects in this gene.

[4] Genetic and biochemical studies suggest the MSH3 proteins are only involved in repair of large loops.

[5] Mutants show an increased rate of small duplications consistent with a possible role in loop repair.

[6] Genetic and biochemical studies suggest that MSH6 proteins are only involved in the repair of base:base mismatches and small loops.

[7] The last two of these may not be true orthologs of the others (see Discussion).

[8] I suggest changing the names of the sequences in this groups to MutS2 to reflect their distinctness from the proteins in the *MutS1* subgroup.

*Information not available.

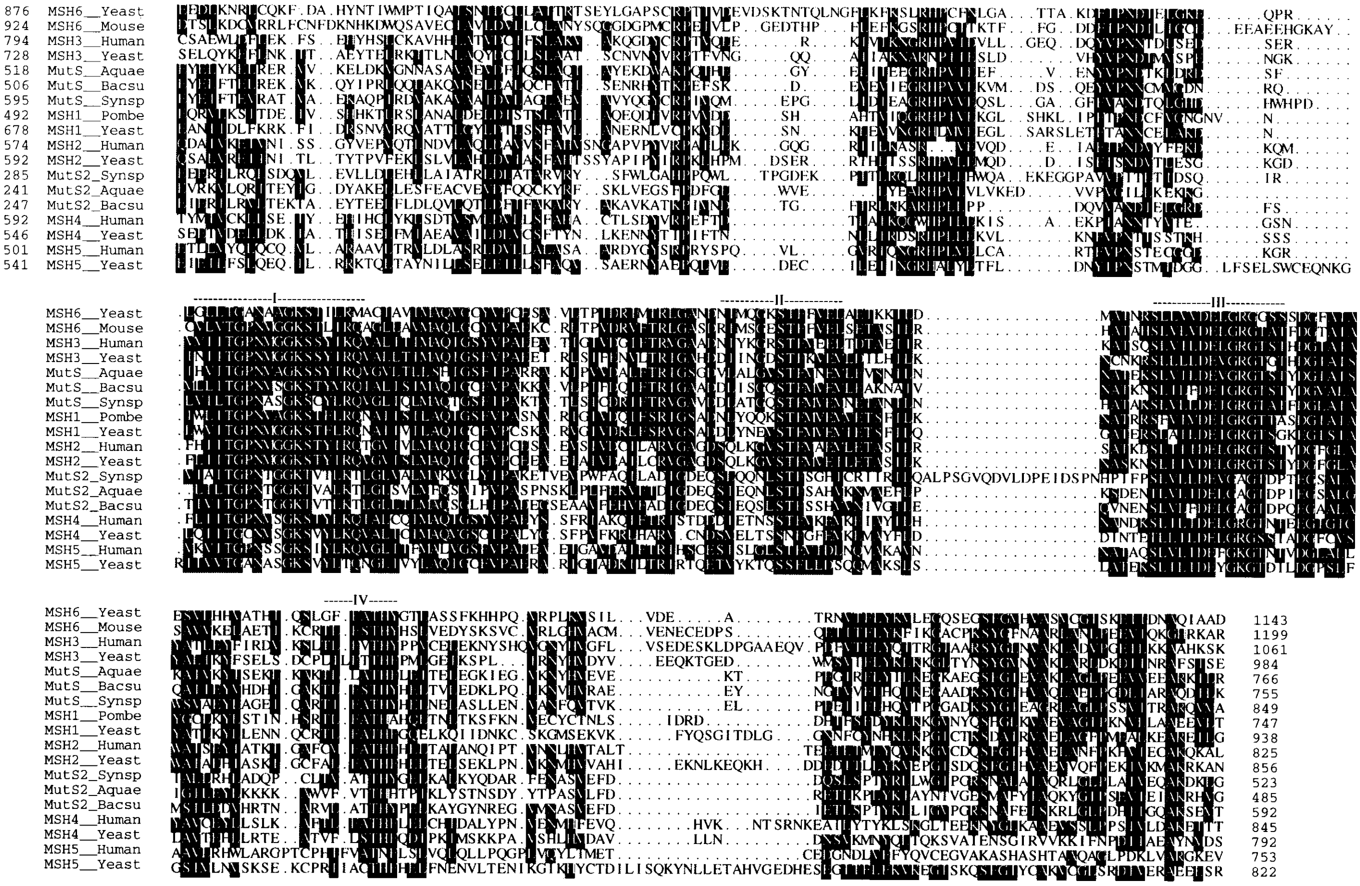


Figure 1. Alignment of a conserved region of the MutS proteins from representative members of the MutS family. The alignment was generated using the *clustalw* and *clustalx* programs and modified slightly manually. Shading was done based on degree of identity or conservation using the MacBoxshade program. Previously described MutS motifs are referred to by roman numerals. The beginning and ending amino acids for each protein are numbered.

events and therefore may not be orthologs of the *MutS2* proteins. Nevertheless, they are close relatives of the *MutS2* subfamily.

Examination of the species represented in each orthologous group can help determine when that group originated. For example, all the eukaryotic subfamilies except *MSH1* include proteins from yeast and humans suggesting that these subfamilies originated prior to the divergence of the common ancestor of fungi and animals. Similarly, the *MutS1* and *MutS2* subfamilies are composed of proteins from diverse bacterial species, including some of the deeper branching bacterial taxa (e.g., *D.radiodurans* and *Aquifex aeolicus*). Therefore the origin of these bacterial subfamilies probably predates the divergence of most of the bacterial phyla. While this type of analysis can help time the origin of the orthologous groups, it does not provide any information about how these groups originated. That is, did the orthologous groups originate by gene duplication or lateral transfer? Many other questions also cannot be answered by the simple division into groups of orthologs. Therefore additional analysis is required.

Unusual distributions of MutS orthologs help identify specific evolutionary events. One way to identify particular evolutionary events in the history of a gene family is to analyze unusual distribution patterns of the different orthologs. Such unusual distributions can be explained either by lateral transfer to the species with an ‘unexpected’ presence of a gene, or by gene loss in the lineages with an unexpected absence of certain genes. These two possibilities can be distinguished by comparing the gene tree to the tree of the species from which these genes come. If an unusual distribution is caused by gene loss, then the gene and species trees should be congruent (as though the species which do not encode a particular gene were just cut out of a larger tree of life). If instead lateral transfer caused an unusual distribution, then the gene and species trees should be incongruent.

Analysis of the distribution of proteins used to be relatively haphazard. However, the availability of complete genome sequences allows for the first time the reliable determination (through sequence analysis) of what genes are present or absent in a species. This of course assumes that homologs can be detected by the sequence analysis methods used. Given the level of conservation among a diverse collection of MutS homologs (Fig. 1), it is likely that most MutS homologs were identified using the search methods described here. A simple identification of homologs in a species does not provide a complete picture of gene presence and absence. It is important to determine presence and absence of specific orthologs. This step is another area in which phylogenetic analysis and genome analysis can be

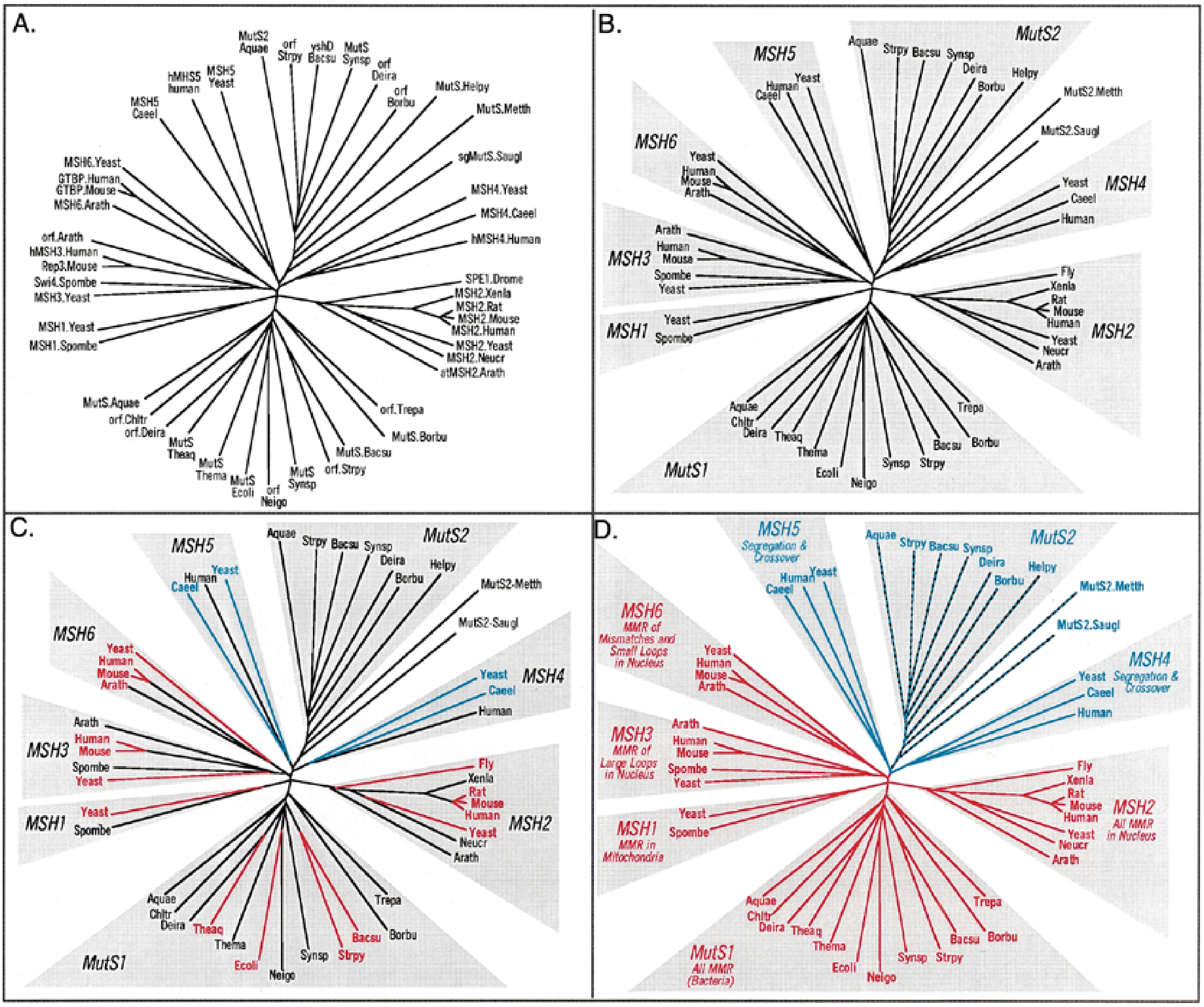


Figure 2. Phylogenomic analysis of the MutS family of proteins. (**A**) Unrooted neighbor-joining tree of the proteins in the MutS family. The tree was generated from a *clustalw* based sequence alignment (with regions of ambiguous alignment excluded) with the PAUP* program. Some of the bacterial MutS1 proteins are left out for clarity. (**B**) Proposed subfamilies of orthologs are highlighted (see Discussion for details). (**C**) Known functions of genes are overlaid onto the tree. For simplicity, only two colors are used, red for mismatch repair and blue for meiotic-crossing over and chromosome segregation. (**D**) Prediction of functions of uncharacterized proteins based on position in the tree.

combined. Although other methods have been developed to determine orthology, phylogenetic methods are preferable (36). Thus, using a combination of sequence searches and phylogenetic analysis, the presence and absence of particular orthologs was determined for all species for which complete genomes are available (Table 3).

Since most of the available complete genome sequences are from bacteria, I focused first on distribution patterns in the bacteria. Every possible pattern of presence and absence of the MutS1 and MutS2 proteins is found in the bacteria (Table 3); some species encode members of both subfamilies, while others encode only one or none. There are two reasonable explanations for this: either rampant gene loss after gene duplication or multiple lateral transfer events. As discussed above, one way of testing which occurred is to compare the phylogenetic trees of the two subfamilies. If there was an ancient duplication, then the branching patterns within the *MutS1* and *MutS2* subfamilies should be identical. However, it is not valid to simply extract the MutS1 and MutS2 evolutionary relationships from the gene tree shown in Figure 2. This is because the MutS1 and MutS2 genes in this tree do not all come from the same species, and species sampling can have a major effect on phylogenetic results (40). To get around this species sampling effect, I generated new trees using only proteins from species that encode both MutS1 and MutS2 (Fig. 3A). As can be seen, the branching patterns in the two subfamilies are congruent when these identical species sets are used. It is important to note that this shared topology is not congruent to that of the rRNA tree of life. The reasons for this are not known but it may simply be due to the limited number of MutS sequences that are available. Regardless, the fact that the

Table 2. Properties of MutS subfamilies

			Boostrap value		
Subfamily	**Conserved Function**	**Comments**	**N.J**	**UPGMA**	**Parsimony**
MutS-I	**Mismatch Repair**				
MutS1	All mismatch repair	In most bacteria.	96	100	25
MSH1	Mitochondrial mismatch repair?	Eukaryotic, not yet found in humans.	100	100	95
MSH2	All mismatch repair in nucleus.	Eukaryotic. Defective in some HNPCC.	100	100	100
MSH3	Repair of loops (small & large) in nucleus.	Eukaryotic. Defective in some HNPCC.	79	100	100
MSH6	Repair of mismatched base pairs & small loops in nucleus.	Eukaryotic. Defective in some HNPCC.	95	100	90
MutS-II	**Chromosome Segregation?**				
MutS2	Unknown	In some bacteria.	74	95	60
MSH4	Facilitate X-over, chromosome segregation	Eukaryotic. Role in humans unknown.	96	97	85
MSH5	Facilitate X-over, chromosome segregation	Eukaryotic. Role in humans unknown.	100	100	55

Table 3. Presence of MutS homologs in complete genome sequences

Species	**# of MutS Homologs**	**Which Subfamilies?**
Bacteria		
Escherichia coli K12	1	*MutS1*
Haemophilus influenzae Rd KW20	1	*MutS1*
Neisseria gonorrhoeae	1	*MutS1*
Helicobacter pylori 26695	1	*MutS2*
Mycoplasma genitalium G-37	0	-
Mycoplasma pneumoniae M129	0	-
Bacillus subtilis 169	2	*MutS1,MutS2*
Streptococcus pyogenes	2	*MutS1,MutS2*
Synechocystis sp. PCC6803	2	*MutS1,MutS2*
Treponema pallidum Nichols	1	*MutS1*
Borrelia burgdorferi B31	2	*MutS1,MutS2*
Aquifex aeolicus	2	*MutS1,MutS2*
Deinococcus radiodurans R1	2	*MutS1,MutS2*
Archaea		
Archaeoglobus fulgidus VC-16, DSM4304	0	-
Methanococcus janasscii DSM 2661	0	-
Methanobacterium thermoautotrophicum ΔH	1	*MutS2*[1]
Eukaryotes		
Saccharomyces cerevisiae	6	*MSH1-6*
Homo sapiens[2]	5	*MSH2-6*

[1] May not be a true ortholog of other members of the *MutS2* subfamily.
[2] Genome not yet complete.

branching patterns of the two subfamilies are congruent indicates that a gene duplication gave rise to these two subfamilies. Thus the absence of *MutS1* and *MutS2* orthologs from some species is most likely caused by gene loss. I inferred likely gene loss events within the *MutS1* and *MutS2* subfamilies by using standard parsimony character state reconstruction (Fig. 3B). The identification of specific gene loss events relies on the accuracy of the species tree onto which the presence and absence of genes is overlaid. The choice of the particular species tree to use is somewhat difficult, since some results suggest that bacterial 'species' do not have a single tree. However, in this case, the choice of the specific tree is not particularly important since all of the inferred gene loss events are in lineages with well-established phylogenies. For example, the inference of gene loss in the mycoplasmas essentially only depends on the well-supported assumption that mycoplasmas are members of the lowGC gram-positive group (since other lowGC gram-positives encode both *MutS1* and *MutS2* orthologs). Thus although the species tree used may not be accurate, the inferred gene loss events are likely to be correct. The implications of specific gene loss events are discussed in more detail below.

The evidence presented above shows that the *MutS1* and *MutS2* subfamilies are most likely related by a gene duplication event. However, the evidence does not specify when this duplication occurred. Based on a variety of evidence, I propose that the duplication was ancient and that the root of the MutS tree is most accurately placed such that it divides the family into two main lineages which I refer to as *MutS-I* and *MutS-II*. *MutS-I* includes the *MutS1*, *MSH1*, *MSH2*, *MSH3* and *MSH6* subfamilies and *MutS-II* includes the *MutS2*, *MSH4* and *MSH5* subfamilies. Three pieces of information support the division into these two main lineages: (i) these two groups were found in all trees regardless of methods or parameters used; (ii) function is generally conserved within but not between lineages (the proteins involved in MMR are all in the MutS-I lineage and those involved in meiotic crossing-over are in the MutS-II lineage) (Table 1); and (iii) such an ancient duplication is consistent with the presence of bacterial and eukaryotic subfamilies in each lineage and is also consistent with the evidence for a duplication prior to the emergence of the major bacterial groups. Since these arguments are somewhat circumstantial and, since the bootstrap values defining the two supergroups are relatively low, this hypothesis should be considered highly tentative. A consensus tree, using the proposed rooting but in which those patterns that are not robust are collapsed, is shown in Figure 4. Even assuming the duplication occurred as proposed, since the relationships among the subfamilies within each lineage are not well resolved in the current analysis, it is not possible to determine the exact patterns of duplications or lateral transfers within each lineage. It is likely that as the sequences of additional members of each subfamily become available the relationships between the subfamilies will become better resolved.

The ancient duplication theory proposed above does not describe all of the unusual distribution patterns in the MutS family. One such pattern is the presence of only one MutS homolog among the three Archaea for which complete genomes are available. This is the MutS2-like protein of *M.thermoautotrophicum*. As discussed above, since the MutS proteins are highly conserved (including the one MutS homolog from Archaea) it is unlikely that other MutS homologs are present in these Archaeal species but were not identified. With the data currently available, it is not possible to resolve the origins of this gene. One reason for this is the lack of a consensus concerning the evolutionary history of the major domains of life. If the Archaea are a sister group to the eukaryotes (as suggested by some studies), then the distribution pattern is probably best explained by gene loss in the history of these Archaea. If instead the bacteria and eukaryotes are sister groups (or even just for the parts of the genome encoding the MutS proteins), then the MutS gene family may have evolved after the Archaea formed a separate lineage. Thus the distribution pattern could be explained simply by lateral transfer to *M.thermoautotrophicum*. Another reason for the difficulty in resolving this unusual distribution

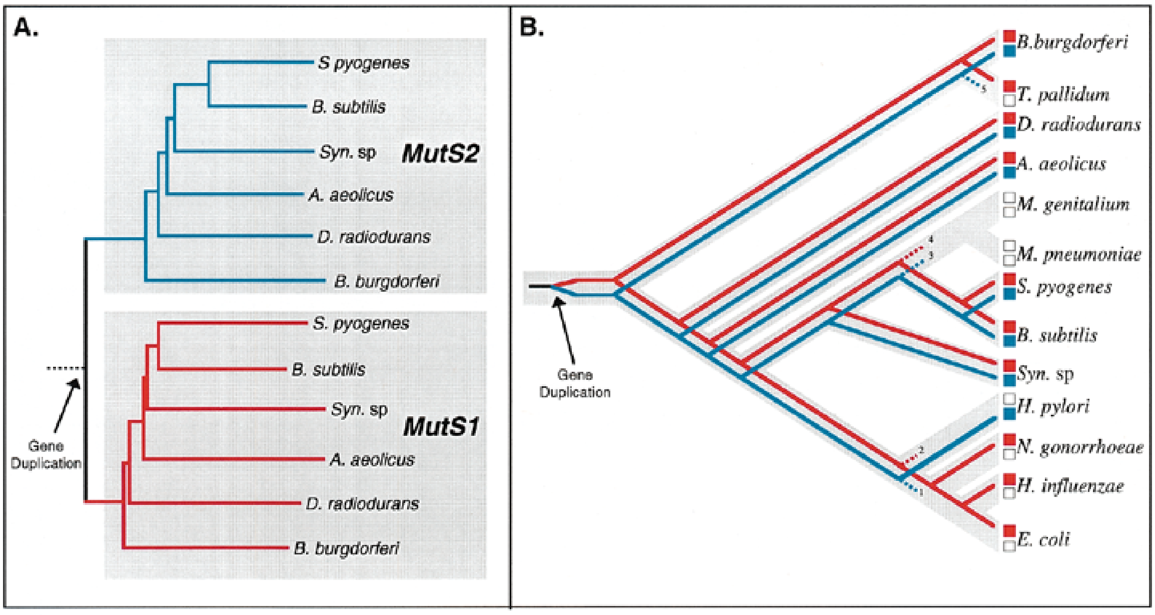


Figure 3. Gene duplication and gene loss in the history of the bacterial MutS homologs. (**A**) Neighbor-joining phylogenetic tree of the *MutS1* and *MutS2* subfamilies (using only those proteins from species with both). The identical topology of the tree in the two subfamilies suggests the occurrence of a duplication prior to the divergence of these bacteria. (**B**) Gene loss within the bacteria. Gene loss was determined by overlaying the presence and absence of MutS1 and MutS2 orthologs onto the tree of the species for which complete genomes are available (since only with a complete genome sequence can one be relatively certain that a gene is absent from a species). The thick gray lines represent the evolutionary history of the species based on a combination of the MutS and rRNA trees for these species. The thin colored lines represent the evolutionary history of the two MutS subfamilies (*MutS1* in red and *MutS2* in blue). Branch lengths do not correspond to evolutionary distance. Gene loss is indicated by a dashed line and each loss is labeled by a number: 1, MutS2 loss in enterobacteria; 2, MutS1 loss in *H.pylori*; 3, MutS2 loss in the mycoplasmas; 4, MutS1 loss in the mycoplasmas; and 5, MutS2 loss in *T.pallidum*.

pattern is that these three species do not represent much of the Archaeal evolutionary diversity. It is likely that additional Archaeal genomes will help resolve the history of the Archaeal MutS homolog(s).

Another unusual distribution pattern is the presence of a MutS homolog (sgMutS) in the mitochondrial genome of the coral *S.glaucum*. Although this mitochondrial genome is not completely sequenced, many other mitochondrial genomes have been and none of these encodes a MutS homolog. In a detailed phylogenetic study, Pont-Kingdon *et al*. found that the sgMutS branched most closely to the yeast MSH1 (39). Since MSH1 is encoded by the nucleus but functions in the mitochondria, this seemed like a possible case of lateral transfer from the mitochondria to the nucleus. However, since the sgMutS did not branch within any bacterial group of proteins and since most mitochondria do not encode a MutS homolog, they concluded that the sgMutS represented a case of 'reverse' lateral transfer from the nucleus to the mitochondria. Although their analysis was sound, it was not complete because they did not include proteins from all of the MutS subfamilies. With the more complete sample of MutS homologs, the sgMutS branches closely to the *MutS2* subfamily and not with the *MSH1* subfamily (Fig. 2). This branching pattern is robust; it was seen in the trees generated by all methods used and it has high bootstrap values. I further tested the robustness of this branch pattern by determining the parsimony score for trees with a variety of lateral transfer scenarios involving the sgMutS and MSH1 proteins including: (i) a mitochondrial origin of the *MSH1* subfamily; (ii) a mitochondrial origin of the sgMutS; and (iii) an *MSH1* origin of the sgMutS (as suggested by Pont-Kingdon *et al*.). Each of these scenarios requires many more steps than the tree in which sgMutS grouped with the *MutS2* subfamily. Thus the results of Pont-Kingdon *et al*. were probably biased by not including proteins from all of the MutS subfamilies. There are two reasonable explanations for the close relationship of the sgMutS to the *MutS2* subfamily. It is possible that there was a lateral transfer of a MutS2-like gene to the mitochondria of an ancestor of *S.glaucum*. Alternatively, the sgMutS may be a true mitochondrial gene and *S.glaucum* may be one of the few species in which this gene still remains. The ability to resolve the origins of the sgMutS will likely improve with the inclusion of more members of the *MSH1* and *MutS2* subfamilies and in particular sequences from alpha-Proteobacterial species which are considered to be the closest living relatives to mitochondria.

Using the evolutionary information

The benefits of using evolutionary analysis in molecular biology come from improving both our understanding of observed molecular characteristics and our ability to make biological useful predictions. What are the particular uses of the evolutionary analysis of the MutS family described above? First, I used the phylogenetic information to infer likely functions for uncharacterized members of the MutS family (Fig. 2b–d). Such a phylogenomic prediction of function is preferable to similarity-based functional predictions for a variety of reasons (see 36 for review). In summary, since function is conserved within orthologous subfamilies, I have assigned predicted functions to uncharacterized genes based on the subfamily in which they are placed. This ortholog rule cannot

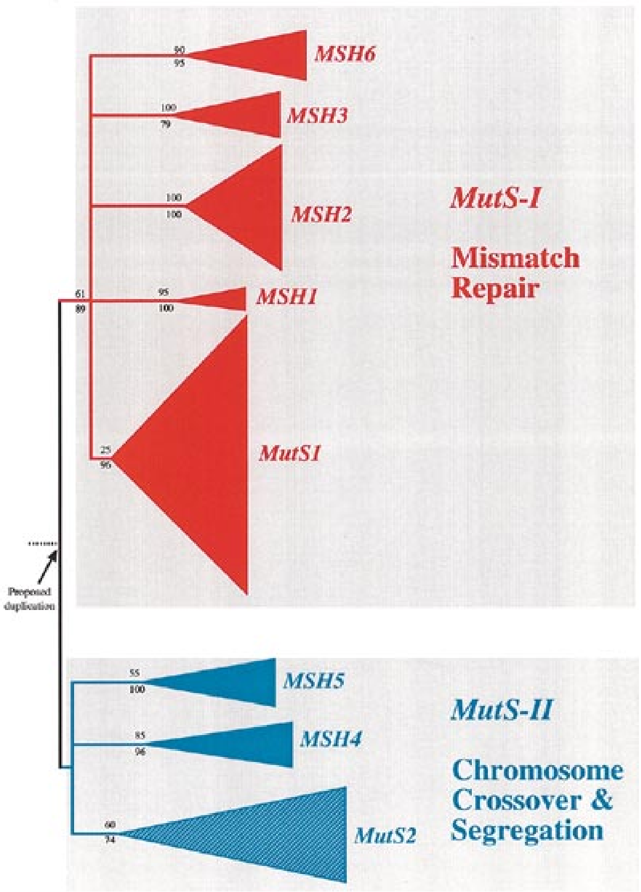


Figure 4. Consensus phylogenetic tree of MutS family of proteins. Branches with low bootstrap values or that were not-identical in trees generated with different methods were collapsed. Only the proposed subfamilies are shown (sequences in each group are listed in Table 1). In addition, two proteins that are related to the *MutS2* subfamily are grouped with it. The height of each subgroup corresponds to the number of sequences in that group and the width corresponds to the longest branch length within the group. Bootstrap values for specific nodes are listed when >40% (neighbor-joining on the top, parsimony on the bottom). The root of the tree was assigned as discussed in the text between the groups labeled *MutS-I* and *MutS-II*. Conserved functions for the different groups are listed.

be applied to those proteins in the *MutS2* subfamily since none of the proteins in this subfamily have a known function. In addition, it cannot be applied to the two MutS2-like proteins since they may not be orthologs of any of the MutS family members. Interestingly, many of the proteins in the *MutS2* subfamily (as well as the two MutS2-like proteins) have been given the name MutS and assigned a likely role in MMR based predominantly on similarity searches (35). The phylogenetic analysis suggests that these functional assignments are likely to be wrong. First, these proteins are all evolutionarily distant from proteins known to be involved in mismatch repair. In addition, many of these proteins are found in species that do not even encode a MutL homolog [e.g. *Helicobacter pylori* (23) and *M.thermoautotrophicum* (41)] and a functional MutL homolog is required for MMR. It is much more reasonable to assign these proteins a possible function in chromosome segregation or crossing-over since they are in the *MutS-II* lineage with proteins in the *MSH4* and *MSH5* subfamilies. Thus the phylogenetic analysis helps suggest what the functions of the genes in the MutS2 subfamily may be and analysis of additional genome data (the presence and absence of MutL homologs) aids in the prediction of function.

The phylogenetic-functional analysis suggests not only that functions have been conserved within orthologous groups but also that the generation of the orthologous groups was accompanied by functional divergence. The evolutionary analysis on its own does not provide a complete explanation of the functions of the MutS genes. There must be some sequence patterns that explain the functional similarities and differences in the family. Since the MutS-family domain is highly conserved among all the MutS-like proteins, this domain is likely to provide some general activity to all the proteins in the family, such as the ability to recognize and bind to unusual double-stranded DNA structures. In addition, there must be some sequence patterns that are conserved within but not between subfamilies (either in these proteins or in regulatory regions) that provide specific functions to each subfamily. The phylogenetic analysis can help identify functionally important motifs because they can be searched for only within subfamilies (42). Thus the phylogenetic analysis can help understand the mechanism of the specificity of each subfamily.

The phylogenetic-functional analysis can be used in combination with gene presence and absence data to predict organismal phenotypes for those species for which complete genomes are available. For example, it is likely that the species that do not encode a protein in the *MutS-I* lineage do not have the MMR process as it has been found in other species. Such an inference is supported by the fact that all species that do not encode a protein in the *MutS-I* lineage also do not encode a MutL homolog (see above and 35). Such a conclusion is supported by the fact that some of the species that do not encode a MutS1 ortholog also have a high mutation rate (e.g. the mycoplasmas) which is consistent with an absence of MMR. However, since it is possible that other enzymatic mechanisms could have evolved to deal with mismatches, without experimental verification it is not possible to know for certain if these species have MMR. Since no function is known for the proteins in the *MutS2* subfamily it is difficult to determine the significance of the absence of orthologs of these genes from species like *E.coli* and *H.influenzae*.

Combining functional predictions for genes with the gene loss analysis allows a better understanding of why the loss of these genes occurred. The gene loss data shows that losses of *MutS1* and *MutS2* occurred in multiple lineages. Many theories have been put forward to explain gene loss during evolution (43,44). Many of these theories involve genome level phenomena such as selection for reduced genome size, or Muller’s ratchet destroying some genes. However, the loss of MutS homologs may be a more gene-specific event, there is likely to be a selective benefit for the loss of MutS genes in some lineages. Defects in MMR have been suggested to be beneficial in certain conditions such as under nutrient stress (45) and selection for pathogenesis (46,47). It is likely that many of these benefits are due to an increased mutation rate, although some may also be due to changes in other functions associated with MMR proteins. While these benefits have been shown by comparing different strains of the same species, it is possible that such benefits may also occur in comparisons between species. For example, it has been suggested that *H.pylori* varies its antigens through a microsatellite mutation process (23). Such mutations would occur at a much higher rate in a MMR deficient strain and could explain the loss of *MutS1* from *H.pylori* sometime in the past.

Conclusions

I have used a combination of phylogenetic reconstruction methods and analysis of complete genome sequences to better understand the MutS family of proteins. Since studies of multigene families and genomes are interdependent it is useful to combine analysis into one study. Phylogenomic methodology similar to that used here can be applied to any multigene family. First, molecular phylogenetic analysis should be used to determine the evolutionary relationships among the genes in the gene family. Then, integration of species information can be used to divide the family into subfamilies of orthologs and to infer evolutionary events such as gene duplications, lateral transfers and gene loss. This evolutionary information can be used in combination with genome information to improve functional predictions for uncharacterized genes. For example, the phylogenetic analysis shows that the proteins in the *MutS2* subfamily are distant and distinct from those involved in mismatch repair, and genome analysis shows that many of the species that encode these genes do not encode other proteins required for mismatch repair. Thus these proteins are likely not involved in mismatch repair. The phylogenomic analysis can also be used to characterize functionally important sequence motifs, to predict the phenotypes of species for which complete genomes are available and to better understand why events such as gene loss and gene duplication may have occurred. In summary, since any comparative biological analysis benefits from evolutionary perspective, the use of evolutionary methods can only serve to improve what can be learned from ever increasing amounts of gene and genome data.

ACKNOWLEDGEMENTS

I would like to thank P. Hanawalt for his continued support; J. Halliday, R. Myers and M.-I. Benito for helpful discussions; F. Taddei, D. Berg, K. Culligan and G. Meyer-Gauen for suggestions on earlier versions of this work; D. Swofford for making PAUP* available prior to release and The Institute for Genomic Research and Oklahoma University for making sequences available prior to publication. The comments of two anonymous reviewers were particularly helpful. I was supported by Outstanding Investigator Grant CA44349 from the National Cancer Institute to P. Hanawalt. I apologize to the authors of primary papers on the various MutS homologs for being unable to cite them all here. References relating to each can be found at the MutS web site: http://www-leland.stanford.edu/~jeisen/MutS/MutS.html

REFERENCES

1 Modrich,P. and Lahue,R. (1996) *Annu. Rev. Biochem*., **65**, 101–133.
2 Kolodner,R.D. (1995) *Trends Biochem. Sci*., **20**, 397–401.
3 Streisinger,G., Okada,Y., Emrich,J., Newton,J., Tsugita,A., Terzaghi,E. and Inouye,M. (1966) *Cold Spring Harbor Symp. Quant. Biol*., **31**, 77–84.
4 Sia,E.A., Jinks-Robertson,S. and Petes,T.D. (1997) *Mutat. Res*., **383**, 61–70.
5 Levinson,G. and Gutman,G.A. (1987) *Nucleic Acids Res*., **15**, 5323–5338.
6 Modrich,P. (1991) *Annu. Rev. Genet*., **25**, 229–253.
7 Parker,B.O. and Marinus,M.G. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 1730–1734.
8 Lynch,H.T., Smyrk,T. and Lynch,J. (1997) *Cancer Genet. Cytogenet*., **93**, 84–99.
9 Kolodner,R. (1996) *Genes Dev*., **10**, 1433–1442.
10 Marsischky,G.T., Filosi,N., Kane,M.F. and Kolodner,R. (1996) *Genes Dev*., **10**, 407–420.
11 Chi,N.W. and Kolodner,R.D. (1994) *J. Biol. Chem*., **269**, 29984–29992.
12 Reenan,R.A. and Kolodner,R.D. (1992) *Genetics*, **132**, 963–973.
13 Reenan,R.A. and Kolodner,R.D. (1992) *Genetics*, **132**, 975–985.
14 Hollingsworth,N.M., Ponte,L. and Halsey,C. (1995) *Genes Dev*., **9**, 1728–1739.
15 Pochart,P., Woltering,D. and Hollingsworth,N.M. (1997) *J. Biol. Chem*., **272**, 30345–30349.
16 Ross-Macdonald,P. and Roeder,G.S. (1994) *Cell*, **79**, 1069–1080.
17 Matic,I., Taddei,F. and Radman,M. (1996) *Trends Microbiol*., **4**, 69–73.
18 Fishel,R. and Wilson,T. (1997) *Curr. Opin. Genet. Dev*., **7**, 105–113.
19 Modrich,P. (1997) *J. Biol. Chem*., **272**, 24727–24730.
20 Culligan,K.M. and Hays,J.B. (1997) *Plant Physiol*., **115**, 833–839.
21 Kunst,A., Ogasawara,N., Moszer,I., Albertini,A., Alloni,G., Azevedo,V., Bertero,M., Bessieres,P., Bolotin,A., Borchert,S., *et al*. (1997) *Nature*, **390**, 249–256.
22 Kaneko,T., Sato,S., Kotani,H., Tanaka,A., Asamizu,E., Nakamura,Y., Miyajima,N., Hirosawa,M., Sugiura,M., Sasamoto,S., *et al*. (1996) *DNA Res*., **3**, 109–136.
23 Tomb,J.F., White,O., Kerlavage,A.R., Clayton,R.A., Sutton,G.G., Fleischmann,R.D., Ketchum,K.A., Klenk,H.P., Gill,S., Dougherty,B.A., *et al*. (1997) *Nature*, **388**, 539–547.
24 Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol*., **215**, 403–410.
25 Altschul,S.F., Madden,T.L., Schaeffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res*., **25**, 3389–3402.
26 Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) *Nucleic Acids Res*., **22**, 4673–4680.
27 Thompson,J. and Jeanmougin,F. (1997). Clustal X Strassborg University, Strassborg.
28 Smith,S.W., Overbeek,R., Woese,C.R., Gilbert,W. and Gillevet,P.M. (1994) *CABIOS*, **10**, 671–675.
29 Eisen,J.A. (1997) In Swindell,S.R. (ed.), *Methods In Molecular Biology, Vol. 70. Sequence Data Analysis Guidebook*. Humana Press Inc., Totowa, NJ, pp. 13–38.
30 Swofford,D. (1991) *Phylogenetic Analysis Using Parsimony (PAUP)* 3.0d. Illinois Natural History Survey, Champaign, IL.
31 Maddison,W.P. and Maddison,D.R. (1992) *MacClade 3*. Sinauer Associates, Inc., Sunderland, MA.
32 Saitou,N. and Nei,M. (1987) *Mol. Biol. Evol*., **4**, 406–425.
33 Felsenstein,J. (1985) *Evolution*, **39**, 783–791.
34 Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A., Merrick,J.M., *et al*. (1995) *Science*, **269**, 496–512.
35 Eisen,J.A., Kaiser,D. and Myers,R.M. (1997) *Nature (Med.)*, **3**, 1076–1078.
36 Eisen,J.A. (1998) *Genome Res*., **8**, 163–167.
37 Pont-Kingdon,G.A., Okada,N.A., Macfarlane,J.L., Beagley,C.T., Wolstenholme,D.R., Cavalier-Smith,T. and Clark-Walker,G.D. (1995) *Nature*, **375**, 109–111.
38 Paquis-Flucklinger,V., Santucci-Darmanin,S., Paul,R., Saunieres,A., Turc-Carel,C. and Desnuelle,C. (1997) *Genomics*, **44**, 188–194.
39 Pont-Kingdon,G.A., Okada,N.A., Macfarlane,J.L., Beagley,C.T., Watkins-Sims,C.D., Cavalier-Smith,T., Clark-Walker,G.D. and Wolstenholme,D.R. (1998) *J. Mol. Evol*., **46**, 419–431.
40 Eisen,J.A. (1995) *J. Mol. Evol*., **41**, 1105–1123.
41 Smith,D.R., Doucette-Stamm,L.A., Deloughery,C., Lee,H., Dubois,J., Aldredge,T., Bashirzadeh,R., Blakely,D., Cook,R., Gilbert,K., *et al*. (1996) *J. Bacteriol*., **179**, 7135–7155.
42 Eisen,J.A., Sweder,K.S. and Hanawalt,P.C. (1995) *Nucleic Acids Res*., **23**, 2715–2723.
43 Maniloff,J. (1996) *Proc. Natl Acad. Sci. USA*, **93**, 10004–10006.
44 Moran,N.A. (1996) *Proc. Natl Acad. Sci. USA*, **93**, 2873–2878.
45 Harris,R.S., Feng,G., Ross,K.J., Sidhu,R., Thulin,C., Longerich,S., Szigety,S.K., Winkler,M.E. and Rosenberg,S.M. (1997) *Genes Dev*., **11**, 2426–2437.
46 LeClerc,J.E., Li,B., Payne,W.L. and Cebula,T.A. (1996) *Science*, **274**, 1208–1211.
47 Sniegowski,P.D., Gerrish,P.J. and Lenski,R.E. (1997) *Nature*, **387**, 703–705.