Nucleic Acids Research, 1995, Vol. 23, No. 14 2715–2723

Evolution of the SNF2 family of proteins: subfamilies with distinct sequences and functions

Jonathan A. Eisen*, Kevin S. Sweder and Philip C. Hanawalt

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

Received March 21, 1995; Accepted June 1, 1995

ABSTRACT

The SNF2 family of proteins includes representatives from a variety of species with roles in cellular processes such as transcriptional regulation (e.g. MOT1, SNF2 and BRM), maintenance of chromosome stability during mitosis (e.g. Iodestar) and various aspects of processing of DNA damage, including nucleotide excision repair (e.g. RAD16 and ERCC6), recombinational pathways (e.g. RAD54) and post-replication daughter strand gap repair (e.g. RAD5). This family also includes many proteins with no known function. To better characterize this family of proteins we have used molecular phylogenetic techniques to infer evolutionary relationships among the family members. We have divided the SNF2 family into multiple subfamilies, each of which represents what we propose to be a functionally and evolutionarily distinct group. We have then used the subfamily structure to predict the functions of some of the uncharacterized proteins in the SNF2 family. We discuss possible implications of this evolutionary analysis on the general properties and evolution of the SNF2 family.

INTRODUCTION

Proteins with extensive amino acid sequence similarity to the yeast transcriptional activator protein SNF2 have been grouped into a protein family. This family includes proteins from a variety of species with roles in cellular processes such as transcriptional regulation, recombination and various types of DNA repair (see Table 1; for reviews see 1,2). In addition to the sequence similarity with other family members, all proteins in the SNF2 family contain sequence motifs similar to those found in many DNA and RNA helicase protein families (1). Proteins with these 'helicase' motifs have been divided into multiple superfamilies based upon amino acid sequence patterns found within the motifs (3). By this method the SNF2 family has been assigned to helicase superfamily 2, which also includes the ERCC3, RAD3, PRIA, EIF4A and PRP16 protein families (3).

The number of proteins assigned to the SNF2 family has increased rapidly over the last few years and continues to expand. Many new family members have been cloned by methods that do not provide any information about their function, such as in genome sequencing projects or by homology-based cloning. Considering the number of proteins in the family, the diversity of their genetic roles and the large number of proteins with unknown function, we thought some insights could be provided by deducing the evolutionary relationships among the family members. Our phylogenetic analysis leads us to propose that the SNF2 family is composed of evolutionarily distinct subfamilies of proteins. We suggest that these subfamilies represent groups of homologous proteins with similar functions and activities and that the functions of some of the uncharacterized members of the SNF2 family can be predicted by their assignment to particular subfamilies. The evolutionary relationships determined here provide insight into the diversity of genetic functions within the family, as well as the likely common biochemical activities of all family members. Finally, we discuss the implications of this analysis for studies of the function of RAD26 and ERCC6 and their role in transcription-coupled repair (TCR) in eukaryotes.

MATERIALS AND METHODS

Sequence alignment

Sequences used in this paper were downloaded from the National Center for Biotechnology Information databases using an electronic mail server (retrieve@ncbi.nlm.nih.gov). Accession numbers are given in Table 1. Similarity searches were conducted using the blastp and tblastn (4), MPsrch (5) and fasta (6,7) computer programs via electronic mail servers (8). Motif searches were conducted using the blocks electronic mail server (9). Alignment of protein sequences was conducted using the clustalv (10) and clustalw (11) multiple sequence alignment programs. The computer-generated alignments were optimized by some minor manual adjustment.

Phylogenetic trees

Phylogenetic trees were generated from the sequence alignments using programs available in the PHYLIP (12), PAUP (13) and GDE (14) computer software packages. Parsimony analysis was conducted using the protpars program in PHYLIP and the heuristic search algorithm of PAUP. Multiple runs searching for the shortest tree were conducted using a variety of starting parameters and branch swapping options. For the distance-based methods we first generated matrices representing the estimated evolutionary distances between all pairs of sequences using the protdist program of PHYLIP, with default settings. Phylogenetic trees were then generated from these matrices using the least-squares method of

* To whom correspondence should be addressed

De Soete (15), as implemented in GDE and the Fitch–Margoliash (16), and neighbor-joining methods (17), as implemented in PHYLIP. Since in both parsimony and distance methods each alignment position (the column containing one amino acid from each species) is assumed to include residues that share a common ancestry among species, regions of ambiguous alignment were excluded from the phylogenetic analysis. For similar reasons regions in which some sequences had alignment gaps were also excluded. Bootstrap re-sampling was conducted by the method of Felsenstein (18), as implemented in PHYLIP. In bootstrapping new data sets are made by re-sampling the alignment positions used in the original data set by random removal and replacement. The result of a single bootstrap is a data set with the same total number of alignment positions as in the original, but in which some original alignment positions may not be represented, while others may be represented multiple times. Phylogenetic trees are generated based upon each of these modified data sets. Comparison of the trees generated with multiple bootstraps can thus give a measure of the consistency of the original tree. We conducted 100 bootstrap replicates for the protpars, neighbor-joining and Fitch–Margoliash methods.

Computer programs

GDE, PHYLIP, clustalv and clustalw were obtained by anonymous FTP from the archive of the Biology Department at the University of Indiana (ftp.bio.indiana.edu). PAUP was obtained from David Swofford (Laboratory of Molecular Systematics, Smithsonian Institution, Washington, DC).

Table 1. Proteins in the SNF2 family

Protein	No. of amino acids	Species	Proposed subfamily	Possible function/comments	GenBank no.	Ref.
SNF2[a]	1703	*S.cerevisiae*	*SNF2*	Transcription activation. DNA-dependent ATPase. Alters chromatin structure?	M61703	58,59
STH1[b]	1359	*S.cerevisiae*	*SNF2*	Cell cycle control. Required for normal growth	M83755	59,60
BRM	1638	*D.melanogaster*	*SNF2*	Transcription activation of homeotic genes	M85049	25,40
BRG1	1022	Mouse	*SNF2*	Binds retinoblastoma protein	S68108	61
BRG1[c]	1613	Human	*SNF2*	Transcription co-activation with hormone receptors	S66910	35,62
hBRM[d]	1586	Human	*SNF2*	Transcription co-activation with hormone receptors	X72889	39,62
SNF2L	976	Human	*SNF2L*	?	M89907	22
ISWI	1027	*D.melanogaster*	*SNF2L*	?	L27127	23
F37A4.8	971	*C.elegans*	*SNF2L*	?	gi458966	63
YB95[e]	1143	*S.cerevisiae*	*SNF2L*	?	Z36114	64
CHD-1[f]	940	Mouse	*CHD1*	Binds DNA	L10410	31
SYGP4	1468	*S.cerevisiae*	*CHD1*	?	gi172808	65
ETL-1	1136	Mouse	*ETL1*	Expressed very early in development. Concentrated in CNS and epithelium	X69942	66
FUN30[g]	1131	*S.cerevisiae*	*ETL1*	Mutants have increased UV resistance	gi171856	67,68
MOT1	1867	*S.cerevisiae*	*MOT1*	Transcription repression. Removes TBP from DNA. DNA-dependent ATPase	M83224	69
RAD26[h]	1085	*S.cerevisiae*	*ERCC6*	Transcription-coupled repair	X81635	32,70
ERCC6	1493	Human	*ERCC6*	Transcription-coupled repair. Defective in Cockayne's syndrome group B	L04791	33
RAD54	898	*S.cerevisiae*	*RAD54*	Recombination repair	M63232	71
DNRPPX	852	*S.pombe*	*RAD54*	?	Z29640	72
YB53[i]	958	*S.cerevisiae*	*RAD54*	?	Z35942	73,74
NUCPRO	1298	Human	*RAD54*	?	L34363	75
NUCPRO	996	Mouse	*RAD54*	?	L34362	75
RAD16	790	*S.cerevisiae*	*RAD16*	Nucleotide excision repair of silent genes	M86929	28,29,76
RAD5[j]	1169	*S.cerevisiae*	*RAD16*	Post-replication repair. GT repeats more stable in mutants	M96644	27,30
RAD8	1133	*S.pombe*	*RAD16*	Mutants have increased sensitivity to UV and gamma irradiation	X74615	26
HIP116A	1009	Human	*RAD16*	DNA-dependent ATPase. Binds HIV and SPH motifs of SV40 enhancer	L34673	77
NPHCG42	506	*A. californica*	None	Viral encoded protein	L22858	78
lodestar	1061	*D.melanogaster*	None	Mutants have excessive chromosome breakage and tangling in mitosis	X62629	79
HepA	968	*E.coli*	None	Induced by DNA damage	M81963	80,81

[a]GAM1, SWI2 and TYE3.
[b]NPS1.
[c]SNF2B.
[d]SNF2A.
[e]YBR245C and YBR1633.
[f]MMKYBP.
[g]YAL019 , YAL001 and YAB9.
[h]GTA1085.
[j]SCTRAAA_3 , YBRO73W and YBR0715.
[h]REV2.

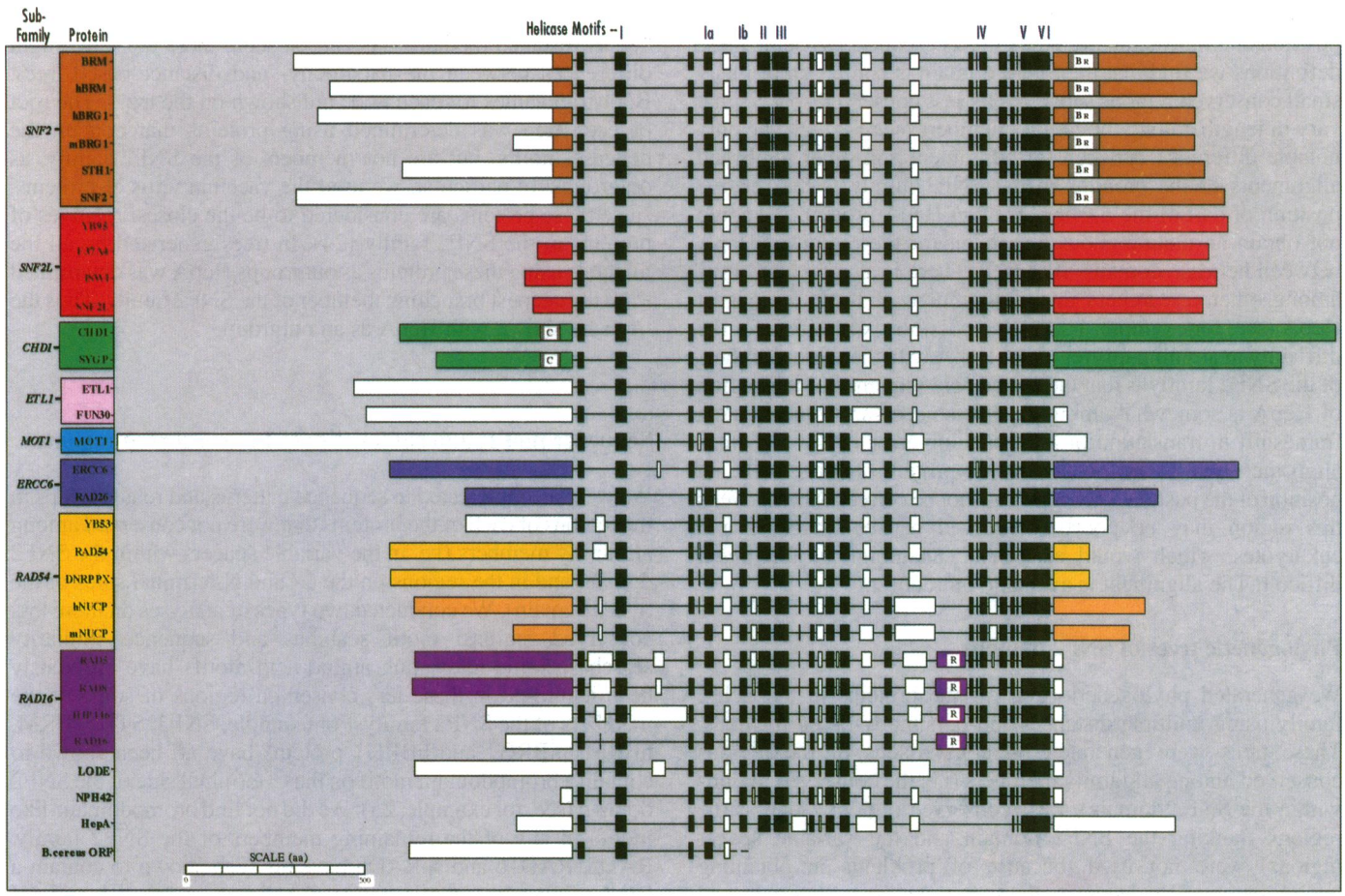

Figure 1. Schematic alignment of the proteins in the SNF2 family. The alignment was generated using the clustalv and clustalw programs and some manual modification. Continuous stretches of amino acids in the alignment are boxed. Alignment gaps are indicated by lines joining boxes. Conserved regions of the SNF2 domain are in black. Colors were chosen to highlight proposed subfamilies. Regions flanking the SNF2 domain are colored for those that show significant similarity to other flanking regions. Blank regions show no significant similarity to other proteins in the family. The presence of motifs is indicated: C = chromodomain, BR = bromodomain, R = RING finger. Scale bar corresponds to numbers of amino acid residues in boxed regions.

RESULTS

Alignment of protein sequences

The presence of a highly conserved domain averaging ~400 amino acids in length has been used to define the SNF2 family (1). We will refer to this conserved region as the SNF2 domain. We first aligned the amino acid sequences of all previously characterized members of the SNF2 family. We then used the SNF2 domains from each of these proteins as query sequences in searches of sequence databases to identify potential additional members of the SNF2 family. A list of all the sequences containing a complete SNF2 domain and some relevant information about these sequences is given in Table 1. In addition to these sequences, we have detected some incompletely sequenced open reading frames that encode peptides that are highly similar to portions of the SNF2 domain. These include a partial open reading frame from chicken (19), two from *Mycoplasma genitalium* (U01723 and U02179 in 20) and many expressed sequence tags from *Caenorhabditis elegans*. The high similarity of the proteins encoded by these sequences to segments of the SNF2 domain suggests that these are also members of the SNF2 family. A new alignment was generated to include all likely members of the SNF2 family. We used this alignment as a block and aligned this block to other proteins with the helicase motifs using the profile alignment method of the clustalv program. A schematic diagram of the alignment of the sequences containing the entire SNF2 domain is shown in Figure 1. A peptide encoded by an incompletely sequenced open reading frame from *Bacillus cereus* is shown in the alignment because it has been previously grouped into the SNF2 family (21). The labeling of particular helicase domains is based on the relative alignment to the suggested helicase domains of these other proteins, as well as previously published assignment of helicase domains to the proteins in the SNF2 family.

The SNF2 domain and the position of the helicase motifs in our final alignment are essentially identical to that presented by others (see, for example, 1,22,23). The degree of amino acid conservation varies greatly within the SNF2 domain. We define conserved regions as those regions in which the alignment is unambiguous, the number of amino acids is the same among all the proteins and the percentage of amino acid similarity between proteins is high. Alignments were considered ambiguous if slight alterations in the alignment parameters, such as changing the scoring matrix used

by the clustalv and clustalw programs, greatly altered the relative position of amino acids from the different sequences. Using these definitions we find that the SNF2 domain is composed of many small conserved regions separated by less conserved spacers that vary in length among the family members (see Fig. 1). The only notable difference between our alignment and other published alignments of the proteins in the SNF2 family is the relative position of part of the *Escherichia coli* HepA protein. We could not obtain an unambiguous alignment for the region of HepA between helicase domains III and V. There is also no consensus among other researchers in the alignment of these regions of HepA (see, for example, 1,21). One possible explanation for the difficulty in aligning this region of HepA with the other members of the SNF2 family is that the amino acid sequence of this region of HepA is somewhat ambiguous. It is necessary to postulate a frameshift in translation or a sequencing error in this region to align the downstream portion of the protein (1) and the exact position of the postulated change may not be correct. Alternatively, this region may be poorly conserved between bacteria and eukaryotes, which would also make unambiguous alignment difficult. The alignment is available on request.

Phylogenetic trees of SNF2 domain

We generated phylogenetic trees of the proteins in the SNF2 family using multiple distance- and parsimony-based methods. These trees were generated by comparisons of the regions conserved among all family members (i.e. the conserved regions within the SNF2 domain). Less conserved regions (such as the regions flanking the SNF2 domain and the variable spacer regions) were not used, because of problems in obtaining unambiguous alignments in these regions (see Materials and Methods) and because there is no established method of scoring alignment gaps in phylogenetic reconstruction. Since the phylogenetic methods are more accurate with more alignment positions, we excluded those proteins, like the *B.cereus* partial sequence, that do not have an entire SNF2 domain. The trees generated using the different distance-based methods were identical in topology. Similarly, the most parsimonious trees found by the two parsimony methods were identical. In Figure 2 we present a comparison of the trees generated by the parsimony versus distance methods. As can be seen, there are only slight differences between the parsimony- and distance-based trees. Bootstrap values for each node are shown on the trees. The root of each tree was determined using proteins that contain the helicase motifs, but are not members of the SNF2 family, as outgroups. In particular, we used the vaccinia virus cI proteins, since these proteins are considered to be the closest relatives of proteins in the SNF2 family (24). In trees generated by all the methods using these proteins as outgroups HepA was determined to be the deepest branching member of the SNF2 family. Thus the trees are shown with HepA as an outgroup.

Sequence motifs and similarities in less conserved regions

We were also interested in sequence patterns and relationships in the regions of each of the proteins that were not conserved among all family members (i.e. in the variable spacers within the SNF2 domain and in the regions on the C- and N-terminal sides of the SNF2 domain). We conducted two types of analyses on these less conserved regions: motif searches and sequence similarity searches. Some interesting amino acid motifs have previously been identified in these less conserved regions of some of the members of the SNF2 family. For example, SNF2, STH1, BRM, hBRM, mBRG1 and hBRG1 proteins have all been shown to contain a bromodomain motif on the C-terminal side of the SNF2 domain (see, for example, 25). We did not find bromodomain-like motifs in any of the remaining members of the SNF2 family. RAD5, RAD16 and spRAD8 have all been shown to contain a RING finger-like motif between helicase motifs III and IV (26–30). We find a similar motif in HIP116A (amino acids 766–836), also between helicase motifs III and IV. Finally, CHD1 has been shown to have a chromodomain motif on the N-terminal side of the SNF2 domain (31). We have found a similar motif in the same relative position in the yeast sequence SYGP4 (amino acids 203–235). No other significant matches to any motif profiles in the blocks database were found. The motifs described above are highlighted in Figure 1.

Table 2. Characteristics of proposed subfamilies

Subfamily	Members with sequence similarity (in region relative to SNF2 domain)			Conserved motifs	Bootstrap values in different phylogenetic methods			Conserved function
	N-terminal side	C-terminal side	Variable spacers		Pars.	Fitch	NJ	
SNF2	All[a]	All		Bromodomain	100	100	100	Transcription activation; remove histones from DNA?
SNF2L	All	All			100	100	100	?
CHD1	All	All		Chromodomain	100	100	100	?
ETL1					97	100	100	?
ERCC6	All	All			100	100	100	Transcription-coupled repair; move stalled RNA polymerase?
RAD16	Some	Some	All	RING finger	47	83	62	
RAD54	Some	Some			81	92	94	Recombination repair?
MOT1	NA	NA	NA		NA	NA	NA	Removes TATA binding protein from DNA

[a]Similarity among all members only over a small stretch of amino acids.

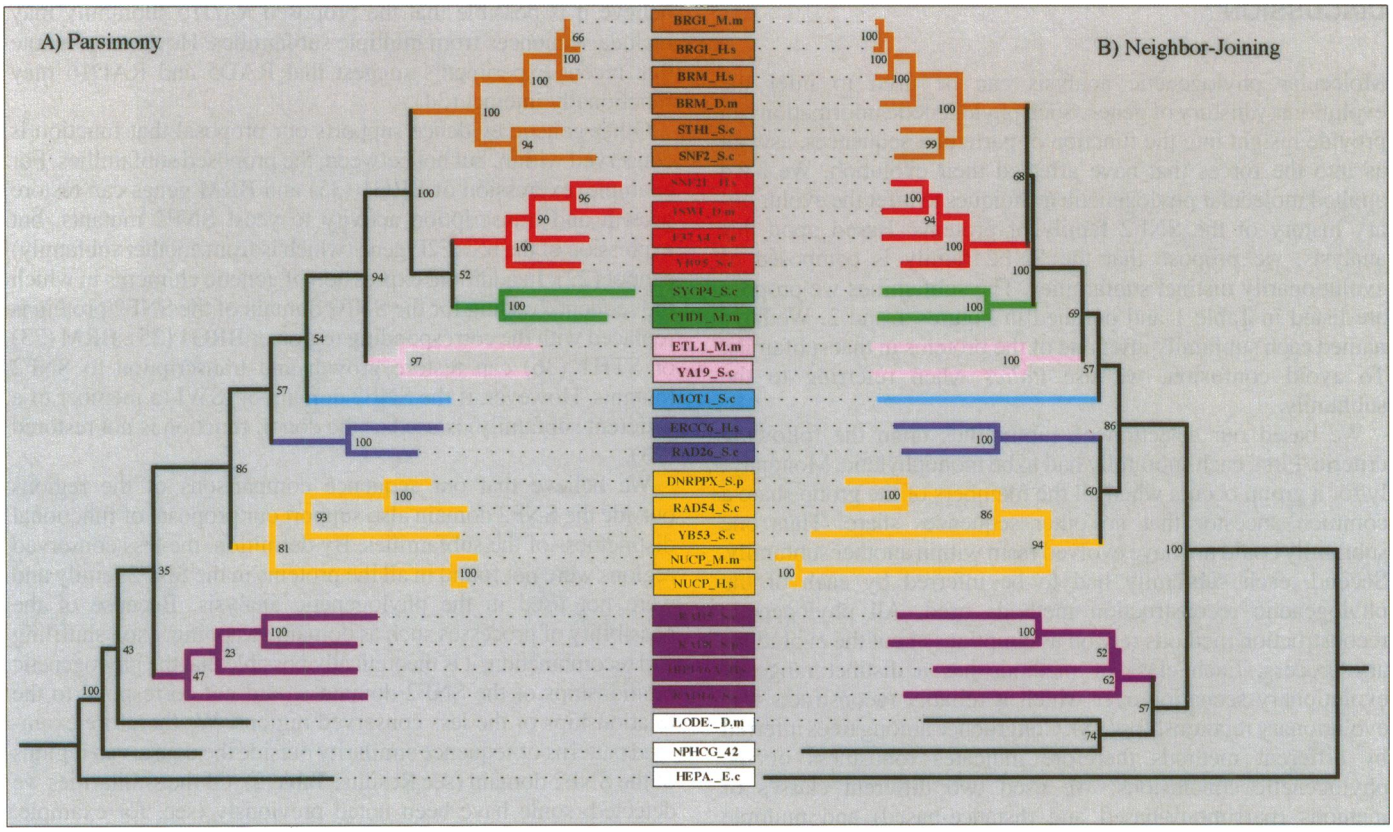

Figure 2. Phylogenetic trees of the SNF2 family of proteins. (**A**) Parsimony tree. (**B**) Neighbor-joining tree. Trees were generated from an alignment generated by the clustalv and clustalw programs. Regions of ambiguous alignment were excluded from the analysis. Bootstrap values, indicated the number of times a particular node was found in trees generated from 100 bootstrap replicates of the alignment are shown on the trees. The roots of the trees were determined by comparisons with other helicase domain containing proteins. Branch lengths correspond to minimum number of inferred amino acid substitutions (A) or estimated evolutionary distance (B). Sequences and branches are colored according to proposed subfamilies. Names are shown in the middle to aid in comparison of the two trees. For more details on tree generation, see Materials and Methods.

We used a variety of sequence comparison programs (see Materials and Methods) to search sequence databases for proteins or possible open reading frames with similarity to the less conserved regions of each of the SNF2 family members. We defined significant similarities as those with $P < 1 \times 10^{-4}$ for at least one of the search methods. Other than in the regions of the motifs described above, the only significant sequence similarities in the less conserved regions of any of the proteins were with other SNF2 family members. In all cases the significant similarities detected were between proteins that branch close to each other in the phylogenetic trees. All similarities detected between two proteins were in comparable regions of the proteins. For example, the regions on the C-terminal sides of the SNF2 domain only showed similarity to other C-terminal regions. Overall, the similarities we found allowed us to divide the SNF2 family into six distinct groups of proteins. All proteins within a group have significant similarity outside the SNF2 domain to all other members of the same group, but not to any other proteins in the SNF2 family. These groups are: 1, SNF2L, ISWI, F37A4 and YB95; 2, CHD1 and SYGP4; 3, ERCC6 and RAD26; 4, hNUCP and mNUCP; 5, RAD54 and DNRPPX; 6, SNF2, STH1, BRM, hBRM, mBRG1 and hBRG1; 7, RAD16, HIP116A, RAD5, spRAD8. MOT1, ETL-1, FUN30, YB53, lodestar, HepA and NPH42 showed no significant similarity outside the SNF2 domain to any other SNF2 family members. A few of the proteins not included in the groups do show small regions of less significant similarity to some other members of the SNF2 family. In all cases these similarities were also between proteins that branch close to each other in the phylogenetic trees.

The regions of significant sequence similarity between group members vary within and among the groups. For example, mBRG1 and hBRG1 are significantly similar throughout their entire lengths, including the regions on the C- and N-terminal sides of the SNF2 domain, as well as the variable spacers. In contrast, mBRG1 and SNF2 show little similarity in the variable spacers, some similarity in the regions on the N-terminal side of the SNF2 domain and extensive similarity in the region on the C-terminal side. To summarize, we have characterized the groups by the regions that are significantly similar among all group members: groups 1, 2, 3 and 4 (both the C- and N-terminal sides of the SNF2 domain); group 5 (N-terminal side); group 6 (C-terminal side with a small region on the N-terminal side); group 7 (the spacer between helicase domains III and IV, which is the location of the RING finger motif in all of these sequences). A summary of the regions of significant sequence similarity is given in Table 2. The regions of similarity among all group members are highlighted in Figure 1.

DISCUSSION

Molecular phylogenetic analysis can be used to infer the evolutionary history of genes. Such phylogenetic information can provide insight into the function of particular sequences, as well as into the forces that have affected their evolution. We have applied molecular phylogenetic techniques to infer the evolutionary history of the SNF2 family of proteins. Based upon this analysis, we propose that the SNF2 family is composed of evolutionarily distinct subfamilies. The subfamilies we propose are listed in Table 1 and outlined in Figures 1 and 2. We have named each subfamily after one of the proteins in that subfamily. To avoid confusion, we use *italics* when referring to the subfamily.

We based our selection of subfamilies upon the following criteria. First, each subfamily had to be monophyletic. Monophyly for a group occurs when all the members of the group share a common ancestor that no other sequences share. Thus one subfamily could not have evolved from within another subfamily. Second, each subfamily had to be inferred by each of the phylogenetic reconstruction methods used. All phylogenetic reconstruction methods rely on assumptions about the evolutionary process. Each class of methods has a distinct range of evolutionary scenarios over which it reliably reconstructs true evolutionary relationships (18). Congruence among trees inferred by different methods therefore indicates robustness of the phylogenetic conclusions. We used two different classes of methods (parsimony-based and distance-based) and multiple types of each method. All proposed subfamilies were found by all methods. Third, the node defining each subfamily had to have high bootstrap values. Bootstrap values for the node defining a subfamily indicate the percentage of times that the sequences in the subfamily grouped together to the exclusion of other sequences in trees generated using different subsamples of a particular alignment. Bootstrapping is thus a method for assessing whether a particular branching pattern has been biased by the sampling of alignment positions. The bootstrap values were very high (between 90 and 100%) for the nodes that define most of the subfamilies (see Table 2). The only proposed subfamily with consistently moderate to low bootstrap values is the *RAD16* subfamily. It is possible that this subfamily would be divided into multiple subfamilies with the availability of sequences from more species.

Our phylogenetic analysis shows that the sequences within the proposed subfamilies are historically more related to each other than to any other characterized proteins, including other members of the SNF2 family. We propose that these evolutionary subdivisions are paralleled by functional subdivisions and therefore that function is conserved within, but not between, subfamilies. In the cases for which the information is available, protein function does appear to be conserved within subfamilies (see Table 1). For example, both members of the *ERCC6* subfamily, RAD26 and ERCC6, are involved in the process of TCR (32,33). In addition, all the proteins in the *SNF2* subfamily for which functional information is available are known to function in transcriptional activation (see Table 1). The *RAD16* subfamily is the only subfamily that includes proteins with known dissimilar genetic functions. This subfamily includes RAD16, which is involved in nucleotide excision repair of non-transcribed regions of the genome, and RAD5 which is involved in post-replication repair and mutagenesis. As discussed above, we believe it is possible that the proposed *RAD16* subfamily may include sequences from multiple subfamilies. However, we note that recent experiments suggest that RAD5 and RAD16 may functionally interact (34).

Other genetic evidence supports our proposal that function is conserved within, but not between, the proposed subfamilies. For example, expression of BRG1 (35) and BRM genes can restore growth and transcription activity to yeast SNF2 mutants, but expression of the hSNF2L gene (which is from another subfamily) cannot (22). In addition, expression of genetic chimeras in which the sequence coding for the SNF2 domain of the SNF2 protein is replaced with the corresponding region of BRG1 (35), BRM (23) or STH1 (36) can restore growth and transcription to SNF2 mutants. However, if the SNF2 domain of ISWI (a member of a different subfamily) is used as the donor, function is not restored (23).

We believe that our sequence comparisons of the regions outside the SNF2 domain also support our proposal of functional distinctness of the subfamilies. By definition, the less conserved regions were not found in all the proteins in the SNF2 family and were not used in the phylogenetic analysis. Because of the possibility of processes such as domain swapping, exon shuffling and recombination, it is theoretically possible that the phylogenetic relationships of the SNF2 domain would not correspond to the relationships of the less conserved regions. We therefore examined patterns of sequence similarity outside the conserved regions of the SNF2 domain (see Results; Table 2). Of the similarities we detected, some have been noted previously (see, for example, 32,35). Most relevant to this study, among SNF2 family members the only significant sequence similarity outside the SNF2 domain is within our proposed subfamilies. In most cases significant similarity outside the SNF2 domain was detected among all members of our proposed subfamilies. This is true for the *SNF2*, *SNF2L*, *ERCC6*, *CHD1* and *RAD16* subfamilies (see Table 2). Thus these regions are conserved within, but not between, subfamilies.

We believe that the sequence conservation within, but not between, subfamilies is due to conservation of function within the subfamilies. The regions conserved within subfamilies may be important in providing specific functions to each of the subfamilies. We believe that analysis of these regions will help identify the function conserved within each subfamily. Some of the proteins in the SNF2 family contain sequence motifs also found in proteins outside the SNF2 family. Other researchers have used the nature of these motifs to help predict the functions of the proteins that have the motifs. We have found that these motifs are conserved within subfamilies and propose that the nature of these motifs may help identify the function conserved within the subfamily. For example, all members of the *SNF2* subfamily contain a bromodomain motif (see Results; Fig. 1). This motif is found in a variety of proteins involved in transcription regulation (25) and it has been suggested that it may be involved in protein–protein interactions (37). It is not known what function the bromodomain provides for the members of the *SNF2* subfamily; it can be deleted from SNF2 (38) and hBRM (39) with no discernible phenotypic effect. Recent studies of BRG1 suggest that the region containing the bromodomain may be involved in binding the retinoblastoma protein (40). Both proteins in the *CHD1* subfamily contain a chromodomain motif. This motif is found in a few other proteins and is proposed to play a role in chromatin compaction (41), but it is not known what role it plays

in the function of CHD1 or SYGP4 (31). Finally, a RING finger motif is found in all the proteins in the *RAD16* subfamily. This motif is related at the sequence and structural levels to the zinc finger motif (42,43). It is found in many proteins that interact with DNA (including the DNA repair protein RAD18, the p53-associated protein MDM2 and the proto-oncogene mel-18) and it is thought that it is involved in DNA binding (42). We believe that the presence of this motif in all the members of the proposed *RAD16* subfamily, but not in any other proteins in the SNF2 family, lends support to the idea that these sequences form a distinct group.

If, as we suggest above, function is conserved within subfamilies, then the functions of some of the uncharacterized proteins in the SNF2 family can be predicted by comparison with other members of the same subfamily. For example, we predict that STH1, the only member of the *SNF2* subfamily for which a genetic role is unknown, is involved in transcription activation, as are all the other members of this subfamily. STH1 is in a monophyletic evolutionary group with the other proteins in the *SNF2* subfamily in every phylogenetic method. In addition, it contains the same sequence motifs, including the bromodomain, found in all the other members of the *SNF2* subfamily. Since STH1 mutants do not have the same phenotype as SNF2 mutants (36), STH1 may have a slightly different function from SNF2. For example, STH1 may be involved in transcription activation only under certain environmental conditions or in certain stages of the cell cycle. We also predict that HIP116A may have some function in DNA repair. HIP116A branches consistently within the *RAD16* subfamily and contains a sequence motif (the RING finger) found in all members of this subfamily, but not in any other members of the SNF2 family. Two of the other members of the *RAD16* subfamily are involved in DNA repair (RAD16 and RAD5) and the third is likely involved in repair (spRAD8) (26). The subfamily structure also allows us to identify likely homologs of uncharacterized mammalian proteins in species in which function may be easier to ascertain. For example, human SNF2L has no known function (22). We suggest that it will be informative to study likely SNF2L homologs, ISWI, YB95 or F37A4, in the more tractable systems of *Drosophila melanogaster*, *S.cerevisiae* and *C.elegans* respectively. Similarly, we believe that elucidation of the function of CHD1 and ETL1 may be facilitated by studying their likely homologs in *S.cerevisiae*, SYGP4 and FUN30, respectively.

The evolutionary relationships among subfamilies are less strongly resolved than those that define the subfamilies. For example, the evolutionary position of some of the subfamilies is different in the parsimony- versus distance-based trees (see Fig. 1). In addition, bootstrap values for the nodes that define the branching patterns between subfamilies are low, indicating that changes in the choice of alignment positions used to generate the trees affect the inferred relationships among subfamilies. More accurate determination of the evolutionary relationships among subfamilies should be possible once more sequences are available in each subfamily. However, we believe that most of the overall topology of the relationships among subfamilies will not change significantly from that presented here. For example, the *SNF2*, *CHD1* and *SNF2L* subfamilies form a coherent supergroup; the bootstrap values for this supergroup are 100 in all trees and the estimated distances (branch lengths) between these subfamilies are low. In addition, we find it intriguing that the proteins known to be involved in DNA repair have deeper branches than those known to be involved in transcription. It is possible that the transcription functions evolved later in the history of this family. However, until more is known about the genetic and biochemical activities of many of the proteins in the SNF2 family, the implications of the inter-subfamily relationships are unclear.

Regardless of the specific phylogenetic relationships among the subfamilies, it is apparent from the number of proteins in the SNF2 family from single species that there have been many duplications in the history of the SNF2 family (see Table 1). We believe the phylogenetic analysis reveals a great deal about the timing of these duplications. Since *S.cerevisiae* has a representative in each subfamily and mammals have a representative in all but the *MOT1* subfamily, we believe that many of the duplications occurred before the separation of fungal and animal ancestors. The rooting of the tree with HepA and the absence of bacterial representatives from the rest of the tree suggests that the majority of the duplications occurred after the separation of bacterial and eukaryotic ancestors. Until complete bacterial genomes are available it is impossible to know for certain if any bacterial species encodes multiple members of the family. Unfortunately, the only likely members of this family from bacterial species other than *E.coli* have not been sequenced completely and are currently too short to use reliably in phylogenetic methods. Complete sequences of these will help better determine the history of these proteins in bacteria. Since in most cases all the proteins within a subfamily contain sequence motifs that are not found in any other members of the SNF2 family, we propose that many of the duplications of the SNF2 domain were accompanied by the addition of these subfamily-specific motifs.

The high conservation of amino acid sequence in the SNF2 domain has led to much speculation about whether any particular biochemical activity is shared by all members of the SNF2 family. The presence of the helicase motifs in the SNF2 domain has been used to suggest that the conserved activity is helicase activity. While helicase activity is needed for the processes (i.e. transcription, recombination and DNA repair) in which these proteins are known to be involved, helicase activity has never been detected in any protein in the SNF2 family. This is despite extensive efforts to detect such activity, especially for SNF2 (44) and MOT1 (45). Despite the presence of the motifs, Henikoff proposed that the SNF2 proteins are not helicases (24) and that the 'helicase' motifs are indicative of a broader DNA-dependent ATPase activity of which helicase activity is a subset. Consistent with this proposal, SNF2, MOT1 and HIP116A have all been shown to be DNA-dependent ATPases. Thus the SNF2 family members may share another activity that requires a DNA-dependent ATPase function.

We believe that the phylogenetic analysis presented here may help understand the common function of the proteins in the SNF2 family. For example, the apparent massive duplication in eukaryotes suggests either that there is something specific about eukaryotes that required or allowed for the diversification of this protein family or that there is something in bacteria that prevented the diversification. Understanding what influenced this diversification in eukaryotes might provide a clue about the common function of these proteins. We believe that recent work on MOT1 helps identify what that eukaryote-specific factor is. Auble *et al.* have shown that MOT1 functions to remove TATA binding protein (TBP) from DNA. They suggest that the common function of the SNF2 family members is the ability to remove proteins from DNA utilizing the energy of ATP hydrolysis (45).

We believe that this activity may have been particularly important during the early evolution of eukaryotes, because of the higher complexity of DNA packaging with proteins and other protein–DNA interactions in eukaryotes versus bacteria. Auble *et al.* suggest that the particular details of protein removal from DNA varies among SNF2 family members. We suggest that these specific details will be conserved within our proposed subfamilies. For example, if the suggestion that SNF2 functions to remove histones from DNA (see, for example, 46) is confirmed, we would suggest that hBRM, BRM, BRG1 and STH1 will have similar activities.

Of the proteins in the SNF2 family, we are particularly interested in the human ERCC6 protein. ERCC6 protein is defective in individuals with Cockayne's syndrome complementation group B (CS-B) (33). Cockayne's syndrome is an autosomal recessive disorder characterized by growth retardation, severe photosensitivity, developmental abnormalities and neural degeneration. Cells from patients with CS-B lack TCR, the preferential repair of DNA damage on the transcribed strand of an actively transcribing gene relative to the non-transcribed strand of the same gene (47,48). It is not known whether the symptoms associated with CS-B are due to their lack of TCR or to another activity of ERCC6 in transcriptional regulation, as has been suggested (49).

Since its discovery in the DHFR gene in hamster cells (50), TCR has been shown to be widespread (48). Mellon and Hanawalt suggested that the mechanism of TCR might involve blockage of transcription by DNA damage and that the recognition of this blockage serves as a signal to the nucleotide excision repair proteins (51). Selby and Sancar subsequently showed that in an *in vitro E.coli* system TCR is an active process requiring a transcription-repair coupling factor (TRCF) and that this TRCF is encoded by the *mfd* gene. They have also shown that the product of the *mfd* gene can remove an *E.coli* RNA polymerase stalled at a DNA lesion (52–54). Selby and Sancar propose that the Mfd protein also serves to recruit the nucleotide excision repair system to that lesion. The Mfd protein, like ERCC6 and RAD26, contains motifs like those found in helicases. As with the proteins in the SNF2 family, despite the presence of the helicase motifs, helicase activity has not been detected in Mfd (55). Although Mfd and ERCC6 both contain helicase motifs, they are not true homologs; ERCC6 (and RAD26) are more closely related to all the other members of the SNF2 family than to Mfd (55). This suggests that perhaps ERCC6/RAD26 and Mfd do not function in a similar way. Despite this complication, there are many similarities between the eukaryotic and prokaryotic processes of TCR. In an *in vitro* eukaryotic system DNA damage in the transcribed strand of an expressed gene is an absolute block to transcription elongation (56). As in *E.coli*, this RNA polymerase complex stalled at the site of DNA damage must then be moved to allow access for repair proteins (56). The moving of a stalled RNA polymerase is similar to the predicted general function of the SNF2 family of proteins; removing proteins from DNA. Thus we predict that ERCC6 and RAD26 function in the moving of stalled RNA polymerase away from the site of DNA damage. If this is true, the lack of homology between Mfd and ERCC6 suggests that eukaryotes and prokaryotes have separately evolved the ability to move a stalled RNA polymerase. It has been suggested that it would be beneficial to eukaryotes for TCR to allow for continued RNA synthesis after DNA repair (because of the amount of energy invested in synthesizing some large RNAs; 57). Thus, unlike in bacteria, eukaryotes may somehow translocate the RNA polymerase, but not remove it.

In conclusion, we believe that molecular phylogenetics is a useful tool in studies of protein families. In the present case we believe molecular phylogenetics has helped to: (i) understand the common properties of the SNF2 family members; (ii) make reasonable predictions of the functions of uncharacterized members of the family; (iii) divide the family into functionally distinct subfamilies; (iv) identify amino acid sequences conserved within, but not between, subfamilies. These regions conserved within subfamilies are probably important in imparting specific functions to the proteins; therefore the characteristics of these regions (e.g. charge, presence of known motifs) may help identify the activity(s) conserved within the subfamilies. The subfamily-specific activities are also determined in part by the characteristics of the highly conserved SNF2 domain; swapping the SNF2 domain leads to functional proteins only when the donor and recipient are from the same subfamily (see above). Related to this, we have identified proteins that do not share any particular motifs outside the SNF2 domain, but which consistently group together in the phylogenetic analysis. Examples of this include the *ETL1* subfamily, in which FUN30 and ETL1 branch together in every analysis but have no significant sequence similarity outside the SNF2 domain, and the *RAD54* subfamily, which includes two subgroups which show no similarity between the groups. The phylogenetic analysis is particularly helpful is these cases.

ACKNOWLEDGEMENTS

We would like to thank J.H.J.Hoeijmakers for helpful comments on earlier versions of this work and Marc Feldman for the use of computer equipment. KSS was supported by a Post-doctoral Training Grant (T32AR07422) from the National Institute of Arthritis and Musculoskeletetal and Skin Diseases to the Department of Dermatology in the Stanford University School of Medicine. JAE was supported by a National Science Foundation Pre-Doctoral Fellowship and a tuition grant from the Marine Biological Laboratory to attend the 1994 Workshop on Molecular Evolution at which much of this research was initiated. This work was also supported by an Outstanding Investigator grant (CA44349) from the National Cancer Institute to PCH.

REFERENCES

1 Bork,P. and Koonin,E.V. (1993) *Nucleic Acids Res.*, **21**, 751–752.
2 Carlson,M. and Laurent,B.C. (1994) *Curr. Opin. Cell. Biol.*, **6**, 396–402.
3 Gorbalenya,A.E. and Koonin,E.V. (1993) *Curr. Opin. Struct. Biol.*, **3**, 419–429.
4 Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
5 Sturrock,S.S. and Collins,J.F. (1993) MPsrch, Version 1.3. Biocomputing Research Unit, Edinburgh, UK.
6 Pearson,W.R. and Lipman,D.J. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 2444–2448.
7 Pearson,W.R. (1990) *Methods Enzymol.*, **183**, 63–98.
8 Henikoff,S. (1993) *Trends Biochem. Sci.*, **18**, 267–268.
9 Henikoff,S. and Henikoff,J. (1994) *Genomics*, **19**, 97–107.
10 Higgins,D., Bleasby,A. and Fuchs,R. (1992) *Comput. Appl. Biosci.*, **8**, 189–191.
11 Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) *Nucleic Acids Res.*, **22**, 4673–4680.
12 Felsenstein,J. (1989) *Cladistics*, **5**, 164–166.

13 Swofford,D. (1991) *Phylogenetic Analysis Using Parsimony (PAUP) Version 3.0d.* Illinois Natural History Survey, Champaign, IL
14 Smith,S.W. (1991) *Genetic Data Environment, Version 2.2a.* Harvard Genome Laboratory, Cambridge, MA.
15 De Soete,G. (1983) *Psychometrika*, **48**, 621–626.
16 Fitch,W.M. and Margoliash, E. (1967) *Science*, **155**, 279–284.
17 Saitou,N. and Nei, M. (1987) *Mol. Biol. Evol.*, **4**, 406–425.
18 Felsenstein,J. (1985) *Evolution*, 39, 783–791.
19 Funahashi,J., Sekido,R., Murai,K., Kamachi,Y. and Kondoh,H. (1993) *Development*, **119**, 433–446.
20 Peterson,S.N., Hu,P.-C., Bott,K.F. and Hutchison, C.A.I. (1993) *J. Bacteriol.*, **175**, 7918–7930.
21 Kolsto,A.B., Bork,P., Kvaloy,K., Lindback,T., Gronstadt,A., Kristensen,T. and Sander,C. (1993) *J. Mol. Biol.*, **230**, 684–688.
22 Okabe,I., Bailey,L.C., Attree,O., Srinivasan,S., Perkel,J.M., Laurent,B.C., Carlson,M., Nelson,D.L. and Nussbaum,R.L. (1992) *Nucleic Acids Res.*, **20**, 4649–4655.
23 Elfring,L.K., Deuring,R., McCallum,C.M., Peterson,C.L. and Tamkun,J.W. (1994) *Mol. Cell. Biol.*, **14**, 2225–2234.
24 Henikoff,S. (1993) *Trends Biochem. Sci.*, **18**, 291–292.
25 Tamkun,J.W., Deuring,R., Scott,M.P., Kissinger,M., Pattatucci,A.M., Kaufman,T.C. and Kennison,J.A. (1992) *Mol. Cell. Biol.*, **12**, 1893–1902.
26 Doe,C.L., Murray,J.M., Shayeghi,M., Hoskins,M., Lehmann,A.R., Carr,A.M. and Watts,F.Z. (1993) *Nucleic Acids Res.*, **21**, 5964–5971.
27 Ahne,F., Baur,M. and Eckardt-Schupp,F. (1992) *Curr. Genet.*, **22**, 277–282.
28 Bang,D.D., Verhage,R., Goosen,N., Brouwer,J. and van de Putte,P. (1992) *Nucleic Acids Res.*, **20**, 3925–3931.
29 Mannhaupt,G., Stucka,R., Ehnle,S., Vetter,I. and Feldmann,H. (1992) *Yeast*, **8**, 385–395.
30 Johnson,R.E., Henderson,S.T., Petes,T.D., Prakash,S., Bankmann,M. and Prakash,L. (1992) *Mol. Cell. Biol.*, **12**, 3807–3818.
31 Delmas,V., Stokes,D.G. and Perry,R.P. (1993) *Proc. Natl. Acad. Sci. USA*, **90**, 2414–2418.
32 Van Gool,A.J., Verhage,R., Swagemakers,S.M.A., van de Putte,P., Brouwer,J., Troelstra,C., Bootsma,D. and Hoeijmakers,J.H.J. (1994) *EMBO J.*, **13**, 5361–5369.
33 Troelstra, C., van Gool,A., de Wit,J., Vermeulen,W., Bootsma,D. and Hoeijmakers,J.H.J. (1992) *Cell*, **71**, 939–953.
34 Glassner,B.J. and Mortimer,R.K. (1994) *Radiat. Res.*, **139**, 24–33.
35 Khavari,P.A., Peterson,C.L., Tamkun,J.W., Mendel,D.B. and Crabtree,G.R. (1993) *Nature*, **366**, 170–174.
36 Laurent,B.C., Yang,X. and Carlson,M. (1992) *Mol. Cell. Biol.*, **12**, 1893–1902.
37 Haynes,S.R., Dollard,C., Winston,F., Beck,S., Trowsdale,J. and Dawid,I.B. (1992) *Nucleic Acids Res.*, **20**, 2603.
38 Laurent,B.C., Treich,I. and Carlson,M. (1993) *Genes Dev.*, **7**, 583–591.
39 Muchardt,C. and Yaniv,M. (1993) *EMBO J.*, **12**, 4279–4290.
40 Dunaief,J.L., Strober,B.E., Guha,S., Khavari,P.A., Ålin,K., Luban,J., Begemann,M., Crabtree,G. and Goff,S.P. (1994) *Cell*, **79**, 119–130.
41 Paro,R. and Hogness,D.S. (1991) *Proc. Natl. Acad. Sci. USA*, **88**, 263–267.
42 Lovering,R., Hanson,I.M., Borden,K.L.B., Martin,S., O'Reilly,N.J., Evan,G.I., Rahman,D., Pappin,D.J.C., Trowsdale,J. and Freemont,P.S. (1993) *Proc. Natl. Acad. Sci. USA*, **90**, 2112–2116.
43 Barlow,P.N., Luisi,B., Milner,A., Elliott,M. and Everett,R. (1994) *J. Mol. Biol.*, **237**, 201–211.
44 Cote,J., Quinn,J., Workman,J.L. and Peterson,C.L. (1994) *Science*, **265**, 53–60.
45 Auble,D.T., Hansen,K.E., Mueller,C.G.F., Lane,W.S., Thorner,J. and Hahn,S. (1994) *Genes Dev.*, **8**, 1920–1934.
46 Wolffe,A.P. (1994) *Curr. Biol.*, **4**, 525–528.
47 Venema,J., van Hoffen,A., Karcagi,V., Natarajan,A.T., van Zeeland,A.A. and Mullenders,L.H. (1991) *Mutat. Res.*, **255**, 123–141.
48 Hanawalt,P.C. and Mellon,I. (1993) *Curr. Biol.*, **3**, 67–69.
49 Bootsma,D. and Hoeijmakers,J.H.J. (1993) *Nature*, **363**, 114–115.
50 Mellon,I., Spivak,G. and Hanawalt,P.C. (1987) *Cell*, **51**, 241–249.
51 Mellon,I. and Hanawalt,P.C. (1989) *Nature*, **342**, 95–98.
52 Selby,C.P., Witkin,E.M. and Sancar,A. (1991) *Proc. Natl. Acad. Sci. USA*, **88**, 11574–11578.
53 Selby,C.P. and Sancar,A. (1991) *Proc. Natl. Acad. Sci. USA*, **88**, 8232–8236.
54 Selby,C.P. and Sancar,A. (1993) *Science*, **260**, 53–58.
55 Selby,C.P. and Sancar,A. (1994) *Microbiol. Rev.*, **58**, 317–329.
56 Donahue,B.A., Yin,S., Taylor,J.-S., Reines,D. and Hanawalt,P.C. (1994) *Proc. Natl. Acad. Sci. USA*, **91**, 8502–8506.
57 Hanawalt,P.C., Donahue,B.A. and Sweder,K.S. (1994) *Curr. Biol.*, **4**, 518–521.
58 Yoshimoto,H. and Yamashita,I. (1991) *Mol. Gen. Genet.*, **228**, 270–280.
59 Laurent,B.C., Treitel,M.A. and Carlson,M. (1991) *Proc. Natl. Acad. Sci. USA*, **88**, 2687–2691.
60 Tsuchiya,E., Uno,M., Kiguchi,A., Masuoka,K., Kanemori,Y., Okabe,S. and Miyakawa,T. (1992) *EMBO J.*, **11**, 4017–4026.
61 Randazzo,F.M., Khavari,P., Crabtree,G., Tamkun,J. and Rossant,J. (1994) *Dev. Biol.*, **161**, 229–242.
62 Chiba,H., Muramatsu,M., Nomoto,A. and Kato,H. (1994) *Nucleic Acids Res.*, **22**, 1815–1820.
63 Wilson,R., Ainscough,R. Anderson,K., Baynes,C., Becks,M. and Bonfield,J. (unpublished) gi458966.
64 Aljinovic,G., Pohl,F.M. and Pohl,T.M. (unpublished) Z36114.
65 Mulligan,J.T., Dietrich,F.S., Hennessey,K.M., Sehl,P., Komp,C., Wei,Y., Taylor,P., Nakahara,K., Roberts,D. and Davis,R.W. (unpublished) gi172808.
66 Soininen,R., Schoor,M., Henseling,U., Tepe,C., Kisters-Woike,B., Rossant,J. and Gossler,A. (1992) *Mech. Dev.*, **39**, 111–123.
67 Kaback,D.B. and Busey,H. (1992) *Yeast*, **8**, 133–145.
68 Barton,A.B. and Kaback,D.B. (1994) *J. Bacteriol.*, **176**, 1872–1880.
69 Davis,J.L., Kunisawa,R. and Thorner,J. (1992) *Yeast*, **8**, 397–408.
70 Huang,M.E., Chuat,J.C. and Galibert,F. (1994) *Biochem. Biophys. Res. Commun.*, **201**, 310–317.
71 Emery,H.S., Schild,D., Kellogg,D.E. and Mortimer,R.K. (1991) *Genes Dev.*, **5**, 1786–1799.
72 Muris,D.F.R., Vreeken,K., Smit,C., Carr,A.M., Broughton,B.C., Lehman,A.R., Lohman,P.H.M. and Pastink,A. (unpublished) Z29640.
73 Steensma,H.Y. and Van der Aart,Q.J.M. (unpublished) Z35942.
74 Van Der Aart,Q.J.M., Barthe,C., Doignon,F., Aigle,M., Crouzet,M. and Steensma,H.Y. (1994) *Yeast*, **10**, 959–964.
75 Gecz,J., Pollard,H., Consalez,G., Villard,L., Stayton,C., Millasseau,P., Khrestchatisky,M. and Fontes,M. (1994) *Hum. Mol. Genet.*, **3**, 39–44.
76 Schild,D., Glassner,B.J., Mortimer,R.K., Carlson,M. and Laurent,B.C. (1992) *Yeast*, **8**, 385–395.
77 Sheridan,P.L., Schorpp,M., Voz,M.L. and Jones,K.A. (unpublished) L34673.
78 Ayers,M., Howard,S., Kuzio,J., Lopez-Ferber,M. and Possee,R. (1994) *Virology*, **202**, 586–605.
79 Girdham,C.H. and Glover,D.M. (1992) *Gene Expression*, **2**, 81–91.
80 Lewis,L.K., Jenkins,M.E. and Mount,D.W. (1992) *J. Bacteriol.*, **174**, 3377–3385.
81 Iwasaki,H., Ishino,Y., Toh,H., Nakata,A. and Shinagawa,H. (1991) *Mol. Gen. Genet.*, **226**, 24–33.