The Evolution of SMC Proteins: Phylogenetic Analysis and Structural Implications

Neville Cobbe and Margarete M. S. Heck

Wellcome Trust Centre for Cell Biology, Institute of Cell and Molecular Biology, University of Edinburgh, United Kingdom

The SMC proteins are found in nearly all living organisms examined, where they play crucial roles in mitotic chromosome dynamics, regulation of gene expression, and DNA repair. We have explored the phylogenetic relationships of SMC proteins from prokaryotes and eukaryotes, as well as their relationship to similar ABC ATPases, using maximum-likelihood analyses. We have also investigated the coevolution of different domains of eukaryotic SMC proteins and attempted to account for the evolutionary patterns we have observed in terms of available structural data. Based on our analyses, we propose that each of the six eukaryotic SMC subfamilies originated through a series of ancient gene duplication events, with the condensins evolving more rapidly than the cohesins. In addition, we show that the SMC5 and SMC6 subfamily members have evolved comparatively rapidly and suggest that these proteins may perform redundant functions in higher eukaryotes. Finally, we propose a possible structure for the SMC5/SMC6 heterodimer based on patterns of coevolution.

Introduction

DNA is compacted in length by a factor of up to 10,000 to form a metaphase chromosome in a typical metazoan cell. How this remarkable feat of packaging is achieved and how the chromosomes are then faithfully segregated to the daughter cells have been fundamental questions for over 100 years, ever since Walther Fleming first described the dynamic behavior of chromosomes during cell division. Within the past decade, a convergence of data from genetic and biochemical approaches has identified the SMC (structural maintenance of chromosomes) proteins as key players in both chromosome condensation and segregation (Cobbe and Heck 2000). These proteins are a highly conserved and ubiquitous family, found in all eukaryotes for which sufficient sequence data are available and in most of the currently sequenced prokaryotes. The function of these proteins in different aspects of chromosomal behavior is also conserved in these organisms, where they play vital roles in the packaging, repair, and segregation of the genetic material.

No more than one smc gene is found in each of the prokaryotes examined, in which the protein apparently forms homodimers, whereas six paralogs are found in eukaryotes, in which they form at least three types of heterodimers. These heterodimers include the SMC1 and SMC3 proteins (part of the "cohesin" complex), the SMC2 and SMC4 proteins (part of the "condensin" complex), and the SMC5/SMC6 heterodimer, which forms part of a complex involved in DNA repair (Hirano 2002). Although no canonical SMC family members have been found in the enterobacteriaceae, similar phenotypes to smc mutants are displayed by mutations affecting mukB in Escherichia coli (Niki et al. 1992). This gene encodes a protein with a structure that is remarkably similar to that of SMC proteins (Melby et al. 1998), despite limited sequence homology. In addition, although the crenarchaeota do not appear to have any SMC or MukB orthologs, they do contain the related Rad50 proteins. Furthermore, although no SMC, MukB, or Rad50 orthologs are apparent in the genome of Rickettsia, it does nonetheless have a related RecN protein. In fact, only four genera for which a complete genome sequence is currently available (Helicobacter, Buchnera, Chlamydia, and Chlamydophila) do not appear to have either an SMC or a related protein. Thus, it appears that SMC proteins have an extremely ancient origin, reflecting their fundamental role in chromosome dynamics.

As shown in figure 1A, each SMC protein is characterized by a large globular domain at each terminus and a third globular domain forming a flexible hinge near the middle of the molecule, separated by extended coiled-coil structures (Saitoh et al. 1994; Melby et al. 1998) with short gaps in the coiled-coils at various positions (Beasley et al. 2002). Although MukB and Rad50 have also been shown to possess hinge regions (Melby et al. 1998; Anderson et al. 2001; Hopfner et al. 2001), the large globular hinge domain of SMC proteins is so far unique in both its sequence and structure (Haering et al. 2002). Recent work has shown that the hinge domain is the principal site at which these molecules dimerize (Haering et al. 2002; Hopfner et al. 2002). The terminal head domains are required for ATP hydrolysis and have been shown to be the region where non-SMC subunits of the condensin and cohesin complexes bind (Anderson et al. 2002; Yoshimura et al. 2002), similar to the binding of MukE and MukF to the head of MukB (Yamazoe et al. 1999) or the binding of Mre11 to Rad50 (Anderson et al. 2001; Hopfner et al. 2001). Although ATP is not required for DNA binding (Kimura and Hirano 1997; Hirano and Hirano 1998; Kimura et al. 1999), it appears necessary for preferential binding to positively supercoiled substrates (Kimura et al. 1999). Likewise, ATP binding (but not hydrolysis) is also required for the enhanced aggregation of the B. subtilis Smc with ssDNA (Hirano and Hirano 1998).

Among the most conserved motifs in all SMC proteins are the N-terminal Walker A (P-loop) and C-terminal Walker B (DA box) ATPase motifs (Saitoh,
SMC proteins share another conserved motif (LSGG) upstream of the Walker B site, referred to as the “signature motif” of the ABC (ATP-binding cassette) ATPase family, which also includes the structurally similar RecN, Rad50, and MukB proteins (Löwe, Cordell, and van den Ent 2001). This motif has been shown to play a pivotal role in the dimerization of Rad50 head domains (Hopfner et al. 2000) and probably serves a similar function in other ABC ATPases such as the SMC, MukB, and RecN proteins. Given the striking level of conservation within SMC proteins and other ABC ATPases with extended coiled-coils, the origin of these proteins and their relationship to one another have been a subject of considerable interest. A number of different phylogenies for SMC proteins have been suggested to date that differ in their method of construction and resultant topology (Melby et al. 1998; Cobbe and Heck 2000; Jones and Sgouros 2001; Soppa 2001; Beasley et al. 2002), so the precise relationships between these and related proteins have been unclear. To resolve this issue, we constructed phylogenetic trees containing all six SMC subfamilies, as well as the related MukB, Rad50, and RecN proteins, from a large data set of protein sequences using the combined criteria of minimum evolution and maximum likelihood. In addition, we used the data generated in constructing these trees to test hypotheses about the structure, function, and evolution of these proteins.

Materials and Methods

Phylogenetic Analysis of SMC Proteins

Sequences were assembled into subfamilies initially based on their BLAST (Altschul et al. 1997) scores, and the whole protein sequences were aligned with ClustalW (Thompson, Higgins, and Gibson 1994) and DIALIGN (Morgenstern et al. 1998). Alignments were then inspected manually with the aid of Mview (Brown, Leroy, and Sander 1998) and adjusted where necessary. Trees containing large numbers of OTUs were constructed using PAL (Drummond and Strimmer 2001) and Weighbor (Bruno, Socci, and Halpern 2000), and smaller maximum-likelihood trees of various subgroups were then calculated by quartet puzzling using the Tree-Puzzle program (Strimmer and von Haeseler 1996). These subgroups included members of the eukaryotic and prokaryotic SMC subfamilies shown in figure 2, subfamilies that branch together or closely related sequences within a subfamily (e.g., the archaeal, proteobacterial, or firmicute SMCs in the eubacterial subfamily). In the case of prokaryotic sequences such as archaeal or firmicute SMCs, the branches in these smaller trees were most easily resolved by removing coiled-coil sequences and calculating distances based on the conserved globular domains.

Maximum parsimony was used to confirm the topology of smaller trees containing closely related sequences within SMC subfamilies, using the tree-searching options in PAUP* (Swofford 1999). However, this method was not used to reconstruct the topology of the whole tree, as it can become inconsistent with unequal rates of evolution (Felsenstein 1978; Kuhner and Felsenstein 1994; Tateno, Takezaki, and Nei 1994), which are observed in the different SMC subfamilies. The topology of the eukaryotic...
FIG. 2.—Phylogenetic tree of SMC proteins. The scale bar denotes the number of accepted substitutions in Whelan-Goldman distance units.
SMCs differed both within subfamilies, when different conserved domains of the protein were used to calculate trees, and between subfamilies, even when the whole protein sequence was used (see Supplementary Material online). Horizontal gene transfer is a comparatively rare event in multicellular eukaryotes, and so it seems more likely that the conflict between subfamilies is the result of additional evolutionary constraints imposed by functional interactions between members of the subfamilies. Consequently, the phylogenetic signal leading to the most probable topology was recovered by calculating the maximum-likelihood tree from concatenated alignments of all the SMC subfamily members for which sequence was available. This widely used approach (Gouy and Li 1989; Baldauf et al. 2000; Hausdorf 2000; Brown et al. 2001; Nei, Xu, and Glazko 2001; Springer et al. 2001; Wolf et al. 2001; Brochier et al. 2002; Lang et al. 2002; Suzuki, Glazko, and Nei 2002) is employed here on the assumption that each of the individual eukaryotic SMC sequences shares the same vertical pattern of inheritance and therefore should follow the same branching order. As these trees using concatenated alignments had bootstrap values of 100% at all nodes (online supplementary figure) and in turn led to the best alignments when used as guide trees, it seems likely that the most probable topology is recovered this way. For those taxa in which gene duplications have occurred within a subfamily, it was found that the separate use of either paralog gave rise to the same topology, after testing all possible combinations of paralogs within subfamilies.

In assembling the complete tree, the phylogenetic information of as many taxa as possible was used to break up long branches (Graybeal 1998; Zwickl and Hillis 2002). Additionally, tests were performed using the less divergent members of subfamilies to reconstruct trees in any cases where long-branch attraction (Lyons-Weiler and Hoelzer 1997) was suspected, by confirming that the same topology existed between other subfamily members. Any ambiguously assigned branches (typically, with a bootstrap probability less than 60%) were compared with all remaining competing tree topologies in which the suspect branch was placed elsewhere, to establish which position contributed least to the total length. The relationship between the subfamilies was established both by calculating maximum-likelihood trees with the least divergent representatives of the different subgroups and by determining the topology yielding the minimum-evolution tree when all sequences were included. Finally, the branch lengths of the overall tree were calculated using the Whelan-Goldman model of amino acid substitution (Whelan and Goldman 2001), with rate heterogeneity among sites modeled according to a gamma distribution with 16 rate categories. The gamma distribution shape parameter was estimated from the data set using the Tree-Puzzle program (Strimmer and von Haeseler 1996).

The phylogenetic tree of coiled-coil containing ABC ATPases was constructed using essentially the same methods as the SMC tree, in which smaller maximum-likelihood trees of each subfamily were calculated using Tree-Puzzle (Strimmer and von Haeseler 1996), and the relationship between the subfamilies was established both by calculating maximum-likelihood trees with the least divergent sequences and by determining the topology yielding the minimum-evolution tree when all sequences were included. Similarly, the final tree of coiled-coil-containing ABC ATPases was constructed by Tree-Puzzle, using the Whelan-Goldman model of amino acid substitution with a gamma distribution. However, branch lengths for this tree were calculated using distances estimated from the conserved globular domains of these proteins, excluding coiled-coils.

Examples of maximum-likelihood trees used to construct the final trees presented in figure 2 and figure 5 are provided as Supplementary Material online, together with confidence bootstrap support values calculated using PAL (Drummond and Strimmer 2001) and Tree-Puzzle (Strimmer and von Haeseler 1996).

Analysis of SMC Codon Usage

The general pattern of codon usage in the genome of different bacteria was compiled using data from the Codon Usage Database (http://www.kazusa.or.jp/codon/). Grantham’s D statistic was calculated using the Wisconsin Package programs CodonFrequency and Correspond version 10.3 (Accelrys Inc) in the following species: Pseudomonas aeruginosa, Mycobacterium tuberculosis, Caulobacter crescentus, Deinococcus radiodurans, Treponema pallidum, Mesorhizobium loti, Thermotoga maritima, Borrelia burgdorferi, Listeria monocytogenes, Neisseria meningitidis, Clostridium acetobutylicum, Bacteroides subtilis, Agrobacterium tumefaciens, Staphylococcus aureus, Xylella fastidiosa, Streptococcus pneumoniae, Aquifex aeolicus, Prochlorococcus marinus, Synechocystis, Nostoc, Thermosynechococcus elongatus, Synechococcus, Archaeoglobus fulgidus, Halobacterium, Pyrococcus furiosus, Methanosarcina acetivorans, and Methanococcus janaschii. The codon bias index and the effective number of codons were calculated using the CodonW program (http://www.molbiol.ox.ac.uk/cu/).

The bias in GC content at silent third codon positions was calculated by developing the bias(GCj) statistic. To correct for biases in GC content caused by a biased amino acid composition in the encoded protein, the GC content at third nucleotide positions was calculated using relative frequencies of codons. The statistic used is defined as follows:

\[
GC_j = \frac{1}{2n} \sum_{j=1}^{31} \sum_{i=1}^{n} \frac{\gamma_j}{\Gamma_j}
\]

where \( n \) is the number of alternative codons ending in either a G or a C for the \( j \)th group of synonymous codons, \( \gamma_j \) is the frequency with which such a codon occurs in the coding sequence, and \( \Gamma_j \) is the sum of the frequencies for each alternative codon in the \( j \)th synonymous group, regardless of GC content. To correct for changes in composition resulting from mutational bias at the same locus, the GC content of the remaining first and second codon positions was calculated (in the same way as for GCj), as well as that of any introns. The arithmetic mean of these three values was taken to be the expected GC content for the locus, and the bias(GCj) statistic for a gene was then calculated as
SMC Evolutionary Rate Correlation

The divergence of each eukaryotic SMC protein region was calculated by summing the maximum-likelihood branch lengths as far as the presumed root where the eukaryotic and prokaryotic branches meet. The extent of correlation between the estimated evolutionary rate in different parts of the proteins was then calculated using the Pearson product-moment correlation coefficient for each pair of SMC protein regions. Similar trends were also found using the Spearman rank correlation coefficient.

Results

Phylogenetic Analysis of SMC Proteins

The final maximum-likelihood tree of 148 SMC sequences is presented in figure 2, with only one representative of each prokaryotic genus displayed. Among the eukaryotic SMC proteins, it was consistently observed that SMC1 and SMC4 branch together (these are the larger SMC subunits of the cohesin and condensin SMC heterodimers, respectively). Similarly, the smaller subunits, SMC3 and SMC2, appear to originate from the same gene duplication event, and SMC5 and SMC6 (which heterodimerize as part of a DNA repair complex [Taylor et al. 2001; Fujioka et al. 2002]) also branch together. All eukaryotes whose genomes have been substantially sequenced, from microsporidia to humans, contain these six subfamilies of SMC proteins, suggesting that the duplication events giving rise to each subfamily must have occurred either before or very soon after the origin of eukaryotes.

It is also noteworthy that the rate of accepted amino acid substitution varies among different eukaryotic taxa within each subfamily. In particular, a number of SMC proteins in Caenorhabditis elegans seem to be exceptionally divergent. However, this can be explained by the observation that the more rapidly evolving SMCs are either those that have arisen from a gene duplication (such as SMC4z and DPY-27) or SMC proteins that form heterodimers with both of the duplicated SMCs (such as MIX-1) (Hagstrom et al. 2002). This is consistent with a correlation between the large number of duplicated genes in the C. elegans genome (Coghlan and Wolfe 2002; Gu et al. 2002) and its rapid evolutionary rate (Aguinaldo et al. 1997; Mushegian et al. 1998). Conversely, the rate of evolution appears relatively slow for the plant and vertebrate SMCs, with the exception of the SMC1β proteins, which are probably less evolutionarily constrained, as they appear to be required only during meiosis in males (Revenkova et al. 2001).

Examining the tree, it appears that the branch lengths between more closely related organisms (such as the vertebrates or diptera) are longer for condensin SMC proteins than for the corresponding cohesins, suggesting either that condensins may be more ancient or may evolve more rapidly. However, the SMC proteins of both complexes appear to be similarly essential for viability in a range of organisms (Strunnikov, Larionov, and Koshland 1993; Saka et al. 1994; Strunnikov, Hogan, and Koshland 1995; Michaelis, Ciosk, and Nasmyth 1997; Lieb et al. 1998; Steffensen et al. 2001; Hagstrom et al. 2002; Liu et al. 2002; Siddiqui et al. 2003), so one might expect their rates of evolution to be broadly similar. Indeed, a significant rate difference is only evident between the more closely-related organisms for which all SMC protein sequences are available, whereas the mean pairwise distances for condensin and cohesin SMCs in more divergent groups seem to approach similar values (fig. 3A). Furthermore, the overall number of substitutions between condensin or cohesin SMC proteins and their presumed ancestor appear to be essentially the same (fig. 3B). It is therefore possible that the trend shown in figure 3A may reflect slightly more rapid evolution in condensin SMC proteins (detectable only by examining the smallest pairwise distances, in which the number of accepted substitutions and speciation events are minimized).

Although condensin SMCs appear to show a higher substitution rate among closely related species than cohesin SMCs, the mean distances within subfamilies of these proteins (averaged across all condensin and cohesin SMCs for each pairwise comparisons between different organisms) are about half (0.54 ± 0.134) the corresponding distances between SMC5 and SMC6 proteins. Similarly, it can be seen in figure 3B that the mean distance of SMC5 and SMC6 to the presumed root with prokaryotes is approximately twice as great as that of other eukaryotic SMC proteins. As discussed later, this elevated sequence divergence may reflect a reduced selection pressure on these proteins compared with the other SMCs, which are required during every cell division, and is thus consistent with the suggested general role for these proteins, specifically in response to DNA damage (Lehmann et al. 1995; Mengiste et al. 1999; Verkade et al. 1999; Fousteri and Lehmann 2000).

Analysis of Prokaryotic SMC Codon Usage

In general, the phylogeny of prokaryotic SMC proteins agrees well with their accepted taxonomic distribution, although the unusually short SMC protein in Borrelia burgdorferi results in this protein clustering with other outgroup species, away from other spirochaetes such as Treponema. However, as previously suggested (Soppa 2001), it would appear from the tree shown in figure 2 that at least two cases of horizontal gene transfer (HGT) have
occurred, resulting in the SMC proteins of the cyanobacteria (*Prochlorococcus marinus, Thermosynechococcus elongatus, Nostoc, Synechocystis, and Synechococcus*) as well as the eubacterial species *Aquifex aeolicus*, most closely resembling those of the archaea. On the other hand, phylogenetic trees based on concatenated alignments of protein or rRNA sequences (Wolf et al. 2001; Brochier et al. 2002) firmly place these species among the remaining eubacteria, as do our own trees based on sequences for 16S rRNA or proteins such as EF-Tu and RecA (data not shown).

To investigate this HGT phenomenon further, codon usage analysis was conducted to examine whether the pattern of codon usage in the cyanobacterial SMC genes more closely resembled that of the archaeal SMCs or of other cyanobacterial genes. The differences between codon frequency tables were computed using Grantham’s *D* statistic (Grantham et al. 1981). This statistic was calculated to measure the difference in codon usage between the SMC of a particular prokaryote and the general pattern of codon usage in its genome. In addition, comparisons were made between genomic codon usage and the codon usage of recA orthologs or EF-Tu genes, which are often employed in phylogenetic tree reconstruction because of the apparent rarity with which they experience HGT (Eisen 1995; Baldauf, Palmer, and Doolittle 1996; Brendel et al. 1997; Lopez, Forterre, and Philippe 1999). However, as shown in figure 4A, the pattern of codon usage for SMC genes in *A. aeolicus*, the cyanobacteria, the remaining eubacteria and the archaea more closely match the general pattern for other loci in the genome than do EF-Tu and RecA, which have higher values of *D*. Therefore, we have found little evidence in favor of HGT for *smc* genes from patterns of codon usage.

As the relative level of sequence similarity between the archaeal and cyanobacterial SMC proteins seems hard to account for in terms of convergent evolution, it would appear that their similarity results from their common ancestry (synapomorphy). This synapomorphy is pre-

---

**Fig. 3.**—Divergence of SMC protein subfamilies. (A) Relative distances (mean pairwise distance within each subfamily divided by the mean distance of all subfamilies) of condensin and cohesin SMC proteins within different taxonomic groups. The difference in relative distances between condensin and cohesin SMC proteins becomes most obvious with progressively more closely related organisms, as shown from left to right. (B) Evolutionary distance between members of different eukaryotic SMC subfamilies and the presumed root with prokaryotic SMC proteins. The mean number of accepted substitutions (in Whelan-Goldman distance units) and standard deviation are displayed for organisms with substantially sequenced genomes in which all SMC protein sequences are available.

**Fig. 4.**—Codon usage in SMC genes. (A) Measures of Grantham’s *D* statistic in comparisons between codon usage for SMC, EF-Tu, and RecA in different prokaryotic species and general codon usage patterns in the same organism. (B) Codon usage in *Drosophila* SMC genes. Values of *bias(GC*) displayed above have all been halved to facilitate comparison with trends in the other statistics. (C) The ratio of nonsynonymous (*K*<sub>s</sub>) to synonymous substitutions (*K*<sub>a</sub>) in SMC genes, based on pairwise comparisons between codons in *D. melanogaster* and *D. pseudoobscura* sequences.
Analysis of SMC Codon Usage in *Drosophila*

The dipteran SMC5 and SMC6 coding sequences seem to have an exceptionally large number of substitutions, even allowing for the observation that a large number of genes in *Drosophila* are known to evolve at an unusually high rate (Sharp and Li 1989; Schmid and Tautz 1997; Fay, Wyckoff, and Wu 2002). The possibility that these genes might be nonessential (Jordan et al. 2002; Yang, Gu, and Li 2003) or that they have been subjected to lower levels of selective pressure (Pál, Papp, and Hurst 2001) was therefore investigated by examining the pattern and usage of codons in the *Drosophila* genes encoding SMC5 and SMC6, in comparison with the codon usage bias of the other four SMCs. In *Drosophila* and other organisms, there is a strong positive correlation between gene expression levels and codon usage bias (Duret and Mouchiroud 1999; Kanaya et al. 2001). However, it has also been suggested that synonymous codon usage in *Drosophila* may be biased to enhance the accuracy of translation, as the frequency of preferred codons is significantly higher at conserved amino acid positions (Akashi 1994). As the number of accepted synonymous substitutions may be constrained by strong codon usage bias, this could explain the apparent correlation between the rates of synonymous and nonsynonymous substitutions in *Drosophila* (Comeron and Kreitman 1998). Therefore, if the higher rates of amino acid replacement in *Drosophila* SMC5 and SMC6 result from reduced selective pressures, this may be reflected by a lack of bias for optimal codons (consistent with reduced translational accuracy or efficiency.)

The extent to which the genes encoding SMC proteins in *Drosophila* demonstrate directional bias in codon usage was first investigated by calculating the codon bias index (CBI) (Bennetzen and Hall 1982). As shown in figure 4B, the CBI for SMC5 or SMC6 is lower than the values for cohesin or condensin SMCs. Moreover, the especially low CBI value for SMC5 is closer to that of the negatively biased *yEst-6* pseudogene (accession number AF526559) of *Drosophila*. The remaining SMC genes all have similar values, with the exception of *Drosophila* SMC3 (CAP), which has a comparatively high level of codon bias. As this gene is found on the X chromosome, whereas the other SMC genes are all autosomal, the increased bias and decreased number of substitutions in *Drosophila* SMC3 is presumably a reflection of the increased selection pressure on X-linked genes in the heterogametic sex (Montgomery, Charlesworth, and Langley 1987). To evaluate whether the reduced bias observed in SMC5 and SMC6 might be simply the result of lower expression levels of DNA repair proteins, the CBI for *Drosophila* Rad50 was also calculated. As this is another gene specifically involved in DNA repair (Gorski and Eeken 2002; Engels et al. 2003; Johnson-Schultz et al. 2003) that also has strong sequence and structural similarities to SMC proteins, one might expect it to be expressed similarly to the SMC5 and SMC6 repair proteins. Interestingly, the CBI for Rad50 is less than most SMCs, which is consistent with a threefold-lower abundance of *rad50* mRNA compared with SMC2 and SMC4 during early *Drosophila* embryogenesis (http://genome.med.yale.edu/LifeCycle/), when most mitotic activity takes place. Nevertheless, the CBI for Rad50 is significantly greater than that of SMC5 and SMC6, suggesting that these proteins may not be required to the same extent as Rad50 to repair DNA damage.

Secondly, the effective number of codons (N_c) (Wright 1990) was also calculated. As N_c can take values from 20 (in the cases of extreme bias) to 61 (when all synonymous codons are used equally), the bias in N_c was evaluated as:

\[
\text{bias}(N_c) = \frac{61 - N_c}{41}
\]

This statistic ranges from 0, for completely random codon usage, to 1, if one codon is exclusively used for each amino acid. These values are also plotted in figure 4B and show that bias(N_c) for SMC5 and particularly for SMC6 is most similar to that of the *yEst-6* pseudogene.

Thirdly, in addition to a correlation between tRNA copy number and favored codons, it has also been described that more highly expressed genes in *Drosophila*
tend to have a high GC content at silent sites such as the third nucleotide position of their codons (Shields et al. 1988; Kanaya et al. 2001). On the other hand, this GC content is often reduced in sequences with weaker selective constraints, such as pseudogenes (Shields et al. 1988) (although this depends on their age [Echols et al. 2002]). The bias in GC content at silent third codon positions was evaluated by calculating the bias(GC3) statistic. As shown in figure 4B, the bias(GC3) for SMC5 and SMC6 is at least 20% smaller than that of most other SMC genes (but only about 11% less than that of Rad50), whereas SMC3 shows the highest bias.

Finally, the selection pressures on Drosophila SMC proteins were examined by estimating the average ratio of nonsynonymous to synonymous substitutions (K_a/K_s) at each codon position. This ratio provides a measure of the intensity of purifying, or negative, selection exerted on genes, which results in decreased rates of amino acid replacement and thus K_a/K_s ratios less than one. The distribution of K_a/K_s varies greatly among different classes of genes, with values ranging from 0.02 to 0.36 and a mean ratio of 0.122 for various Drosophila genes (Li 1997). On the other hand, the rates of synonymous and nonsynonymous substitutions are approximately equal if most amino acid variation is neutral, as observed in many pseudogenes (Li, Gojobori, and Nei 1981; Bustamante, Nielsen, and Hartl 2002), whereas K_a/K_s ratios greater than 1 usually indicate adaptive or positive selection (Messier and Stewart 1997; Hughes and Yeager 1998). As shown in figure 4C, the K_a/K_s ratios for SMC2, SMC4, and Rad50 appear to be close to the previously described mean value in Drosophila (Li 1997), whereas SMC1 and SMC3 have considerably lower K_a/K_s ratios. This increased negative selection on SMC1 and SMC3 is consistent with the differences in amino acid substitution rates between cohesin and condensin SMCs in closely related organisms, as described earlier. By contrast, SMC5 and SMC6 display K_a/K_s ratios approximately twice as great as other SMC proteins. In particular, the ratio for SMC6 is similar to the highest K_a/K_s values previously described in Drosophila for chorion proteins (Li 1997). Based on these combined results, it would appear that the SMC5 and SMC6 proteins in Drosophila are under severely reduced selection for preferred codons in comparison with other SMCs and experience dramatically decreased negative selection, which could explain their higher amino acid substitution rate in terms of fewer selective constraints on their evolution in general.

Phylogenetic Analysis of SMC Related Proteins

To reveal how closely related SMC proteins are to similar ABC ATPases, a maximum-likelihood tree was constructed using sequences from these protein families, as shown in figure 5. It is clear from this tree that the closest relatives to the SMC proteins are the archaeal Rad50 proteins, followed by eukaryotic Rad50 and eubacterial SbcC proteins. Many of these Rad50 superfamily proteins have the conserved N-terminal FKS (or FRS) motif (located before the Walker A site), which is present in most of the SMC proteins. However, this motif is absent in the SMC5 and SMC6 subfamily members (in which only the phenylalanine is conserved). The FKS motif is also absent in the MukB proteins, which appear to possess universally conserved NWN and FART motifs instead. In addition, although it was previously suggested that all SMC and MukB proteins uniquely share a conserved N-terminal QG motif (Melby et al. 1998), this dipeptide was found to occur at different locations in SMC and MukB sequences in the alignments generated in this study. Instead, this motif was found at a conserved position in SMC and Rad50 proteins, in agreement with previously published structure-based alignments (van den Ent et al. 1999). Therefore, although it was previously proposed that MukB should be included among the SMC proteins as they adopt a similar structure (Melby et al. 1998), the Rad50 proteins seem to be more closely related at the sequence level. The high level of sequence conservation between MukB proteins argues against the notion that they are a highly divergent group of SMC proteins. Instead, it is possible that their structural similarity may be partly caused by convergent evolution, which is consistent with the similar functions of MukB and SMC proteins in prokaryotes (Cobbe and Heck 2000).

Correlated Rates of Evolution in Different Regions of SMC Proteins

Contrary to previous reports (Melby et al. 1998), we observed that eukaryotic SMC trees reconstructed from individual conserved hinge or N-terminal and C-terminal globular domains often differed both from each other and from phylogenies based on the whole protein sequence. For this reason, we attempted to dissect the different selective constraints on the sequence divergence of these regions by examining how they might coevolve. As interacting proteins tend to evolve at similar rates (Fraser et al. 2002), it was expected that regions of SMC proteins that interact with each other would also coevolve and show correlated rates of substitution.

Correlation coefficients for different regions of SMC proteins were calculated after sorting the maximum-likelihood distances according to the functional complexes within which different subfamily members are found. The Pearson correlation coefficient for each pair of SMC protein regions is plotted in figure 6, together with Spearman’s coefficient of rank correlation. As shown in figure 6B, a clear positive correlation was observed between the evolutionary rates of the β-strands that constitute the interface of the N-termini and C-termini in the same molecule (Löwe, Cordell, and, van den Ent 2001). However, no significant positive correlation was found between the corresponding N-terminal and C-terminal regions of any SMC and its partner (fig. 6A), as would have been expected if the heterodimer existed naturally in an anti-parallel intermolecular conformation. This strongly suggests that the intramolecular structure (fig. 1B) indicated by available in vitro data is also the structure found in vivo. By contrast, a strong positive correlation was observed between the evolutionary rates of the exposed outer portions of the N-terminal and C-terminal domains in SMC2 and SMC4 but not in the corresponding regions.
FIG. 5.—Phylogenetic tree of SMC proteins and other coiled-coil ABC ATPases. The scale bar denotes the number of accepted substitutions in Whelan-Goldman distance units.

[Diagram showing the phylogenetic tree of SMC proteins and other coiled-coil ABC ATPases, with species names and distances indicated.]
of the same molecule (fig. 6C and D). These stronger intermolecular correlations are nonetheless consistent with observed associations between the termini of SMC proteins (Melby et al. 1998; Anderson et al. 2002). Moreover, a significant correlation at the 1% level was observed for intermolecular associations between the coiled-coils of the condensin SMC proteins (fig. 6E). This elevated intermolecular correlation appears to reflect differences between the conformation of vertebrate condensin and cohesin SMC proteins previously observed by rotary shadowing, whereby the coiled-coils of condensin SMCs are intimately associated along most of their length but those of the cohesin complex are well separated (Anderson et al. 2002).

Interestingly, the SMC5 and SMC6 proteins appear to show similarly weak associations between their arms (fig. 6C and E) as SMC1 and SMC3. This suggests that the SMC proteins involved specifically in DNA repair may adopt a similarly open conformation to that of the cohesin SMC proteins (Haering et al. 2002). By contrast, although a significant intermolecular correlation was not found consistently for the rate at which the hinge domains of cohesin SMC proteins evolve, the SMC5 and SMC6 proteins nonetheless showed intermolecular rate correla-
tions greater than 60% at the dimer interface of their hinge domains. This may possibly indicate that SMC5 and SMC6 have a stronger hinge association than other SMC proteins such as SMC1 and SMC3. Although an even stronger positive correlation can be seen for the dimer interface of condensin hinge domains using the Pearson correlation coefficient, this particular result is not supported by nonparametric analyses.

Discussion
The Phylogeny of SMC Proteins

We have investigated the evolutionary radiation of SMC proteins and determined that the closest relatives of these proteins are members of the Rad50 superfamily, as revealed by phylogenetic analysis and the conservation of particular residues. Apart from the suspected cases of HGT described in this study, the tree topology described here generally agrees with the currently accepted taxonomic distribution of the organisms whose SMC proteins have been examined. Interestingly, our consensus tree shows that plant and animal SMC proteins cluster together with fungi as the outgroup, whereas other recent studies based on 18S rRNA and a panoply of proteins have reported that fungi and animals are most closely related to each other (Baldauf and Palmer 1993; Baldauf et al. 2000; Van de Peer et al. 2000). However, the resolution of this trichotomy has been a longstanding controversy, with different studies suggesting that animals are closer relatives to either plants or fungi (Gupta 1995; Wang, Kumar, and Hedges 1999). The tree presented here is more consistent with previous analyses using ribosomal proteins (Veuthey and Bittar 1998) and other concatenated sequences (Gouy and Li 1989), which place plants closer to animals. Alternatively, the phylogeny revealed in the SMC tree may reflect higher rates of substitution in the fungal proteins because of reduced constraints in their interactions with other proteins, as has been suggested in the case of histones H3 and H4 (Thatcher and Gorovsky 1994). For example, at least one protein (AKAP95) with no clear orthologs in other animals, fungi, or plants is thought to interact with SMC proteins in vertebrates (Collas, Le Guelllec, and Tasken 1999; Steen et al. 2000), which may influence the relatively slow rates of SMC evolution in these animals. In the future, it will be interesting to see how the use of SMC sequences, in combination with larger data sets of similarly conserved protein sequences (Brown et al. 2001), might contribute to resolving the relationships among these eukaryotic kingdoms.

Evolution of Eukaryotic SMC Subfamilies

According to the tree in figure 2, it appears that a symmetric duplication of genes encoding the larger and smaller eukaryotic SMCs gave rise to both cohesin and condensin SMC proteins in all eukaryotes. The relatively short branch lengths connecting either the roots of the SMC1/SMC4 or the SMC2/SMC3 lineages to the prokaryotic SMC root also suggest that the first duplication event, giving rise to the primordial eukaryotic SMC heterodimer, occurred very early in eukaryotic evolution. However, it seems that the first SMC heterodimer to arise in eukaryotes may have had more functions in common with the condensin heterodimer, based on the assertion that the primary phenotype in a prokaryotic *snc* null mutant is probably a condensation defect (Britton and Grossman 1999) and the observation that DNA reannealing (Sutani and Yanagida 1997; Hirano and Hirano 1998) and supercoiling (Kimura and Hirano 1997; Lindow, Britton, and Grossman 2002) activities are common features of these proteins but apparently not of cohesins in general (Jessberger et al. 1996; Losada and Hirano 2001; Sakai et al. 2003). Given these functional similarities between condensin and prokaryotic SMCs, it might seem odd that the primary sequences of SMC2 and SMC4 do not appear to be more closely related to prokaryotic SMC proteins than the functionally distinct SMC1 and SMC3 proteins (fig. 3B). However, if condensin SMCs had evolved slightly more rapidly since the divergence of eukaryotes, then this could account for their lack of greater sequence identity with the functionally related prokaryotic SMC proteins. Indeed, a higher substitution rate among condensin SMC proteins than cohesins can be seen in pairwise comparisons between sequences from more closely related organisms, such as vertebrate or dipteran SMCs. This suggests that SMC2 and SMC4 may evolve more rapidly than the cohesin SMCs but that the difference in rate is small enough to be masked by mutational saturation in comparisons between more distantly related species. If so, it is anticipated that additional SMC protein sequences from closely related organisms in other phyla might also reveal more accepted substitutions in pairwise comparisons between condensin SMC proteins than in cohesins. One might speculate that the later evolution of SMC2 and SMC4 involved some level of positive selection for proteins with enhanced condensing activities, given the larger size of eukaryotic chromosomes in general (Bendich and Drlca 2000), whereas SMC1 and SMC3 were recruited to specific roles in maintaining sister chromatid cohesion. Although speculative, a prediction of this hypothesis could be tested by directly comparing the supercoiling activities of condensin and prokaryotic SMCs to see if the condensing activities of the former are indeed more efficient. Moreover, as additional SMC sequences from closely related species become available, this may facilitate the detection of positive selection at particular amino acid sites of condensin SMC proteins by maximum-likelihood approaches (Anisimova, Bielawski, and Yang 2001).

Whereas, differences in the substitution rate of cohesin and condensin SMC proteins become more pronounced in comparisons between closely related organisms, SMC5 and SMC6 consistently show pairwise distances about twice that of the average for the other SMCs. Based on the pattern of codon usage and \(K_s/K_a\) ratios in *Drosophila*, it seems that the large numbers of substitutions in the SMC5 and SMC6 proteins are best explained by an elevated rate of evolution, rather than a more ancient origin. Similar codon usage analyses in other organisms failed to clearly show the same pattern, often because of less pronounced codon bias patterns or
the complicating influence of SMC gene duplications on evolutionary rate. However, as SMC5 and SMC6 are only known to be essential for proliferation in yeasts (Lehmann et al. 1995; Verkade et al. 1999) but not in Arabidopsis (Mengiste et al. 1999), it seems likely that their reduced sequence conservation may reflect reduced selective pressures on these proteins in other species as well. Moreover, as some of the properties of SMC5 and SMC6 (Taylor et al. 2001) are mirrored by the association of SMC1 (or SMC1β in vertebrates) and SMC3 with male meiotic chromosomes (Klein et al. 1999; Eijpe et al. 2000; Revenkova et al. 2001) and the role of SMC1/SMC3 in recombination repair (Jessler et al. 1996; Sjögren and Nasmyth 2001; Kim, Xu, and Kastan 2002; Yazdi et al. 2002), it is possible that these proteins have evolved more rapidly since their function can be partially complemented by the cohesin SMCs. Therefore, the long branch lengths within this subfamily seem to better reflect their less-constrained substitution rate, rather than their greater antiquity as previously suggested (Jones and Sgouros 2001). Although the other coiled-coil ABC ATPases appear to place the root of SMC proteins in the SMC5/SMC6 branch, this does not necessarily imply that these are the ancestral SMC proteins. Instead, it seems more likely that the placement of the root here in the maximum-likelihood tree reflects the average similarity of other coiled-coil ABC ATPases to SMC proteins, regardless of how the SMCs have diverged from each other.

The Conformations of SMC Heterodimers

In agreement with the results obtained previously by in vitro analyses (Hirano et al. 2001; Haering et al. 2002; Hirano and Hirano 2002), our statistical analysis of coevolution in eukaryotic SMC heterodimers strongly favors the intramolecular model of dimer formation (fig. 6A and B). We believe our findings are complementary to the available in vitro data as the observed correlations between substitution rates would be consistent with formation of such intramolecular heterodimers in vivo.

The low correlations between the evolutionary rates at the coiled-coils and termini of SMC5/SMC6 heterodimers suggest that their conformation may more closely resemble cohesin SMC proteins, consistent with their overlapping functions in DNA repair and meiosis. The similarly weak intermolecular correlations for the arms and termini of cohesin SMCs may represent a weaker association between these regions (fig. 7), facilitating the dissolution of sister chromatid cohesion mediated by a proposed annular cohesin complex (Gruber, Haering, and Nasmyth 2003). The weaker association between the arms suggested by our analysis is consistent with the appearance of both the SMC1/SMC3 heterodimer and the cohesin complex when observed by electron microscopy (Anderson et al. 2002; Haering et al. 2002). By contrast, the strong intermolecular correlation for the coiled-coils and the termini of condensin SMC proteins are consistent both with electron microscope images of these proteins (Anderson et al. 2002; Yoshimura et al. 2002) and the high specificity of their coiled-coil interactions (Hirano and Hirano 2002). In the future it will be interesting to see if SMC5 and SMC6 also adopt a similarly open conformation to that of the cohesin SMC proteins when viewed by microscopy. On the other hand, if SMC5 and SMC6 prove to have a stronger hinge association than other SMC proteins, this may reflect a role for this domain in tethering such SMC proteins bound to separate DNA strands during the repair process, as has been suggested similarly for Rad50 (de Jager et al. 2001; Hopfner et al. 2002).

Supplementary Material

Maximum-likelihood trees of selected SMC proteins from eukaryotes in which all SMC protein sequences are available, are presented in an online supplementary figure. Separate trees of the six different subfamilies are shown together with a consensus tree constructed from concatenated alignments of the other sequences (SMC1 to SMC6). Scale bars denote the number of accepted substitutions in Whelan-Goldman distance units and bootstrap percentage support for the nodes are indicated.

Web sites where additional data are available include http://www.kazusa.or.jp/codon/ Codon Usage Database, http://www.molbiol.ox.ac.uk/cu/ CodonW Program, http://genome.med.yale.edu/Lifecycle/Drosophila Gene Expression Levels, and http://www.hgsc.bcm.tmc.edu/blast/?organism=Dpseudoobscura D. pseudoobscura Genome

Acknowledgments

We would like to thank Andrew Lloyd of INCBI (Irish National Centre for Bioinformatics) and the UK Human Genome Mapping Project Resource Centre for use
of computing facilities. In addition, we are grateful to Dietlind Gerloff, Paul McLaughlin, Ken Wolfe, and an anonymous referee for helpful discussions and critical reading of the manuscript. M.H. is a Senior Research Fellow in the Basic Biomedical Sciences, funded by the Wellcome Trust. N.C. has been supported by a Darwin Trust Prize Studentship.

Literature Cited


Gorski, M. M., and J. C. J. Eeken. 2002. The Drosophila rad50 mutants are pupal lethal, however the third instar larvae show elevated levels of anaphase bridges in dividing cells. A. Dros. Res. Conf. 43:201CC.


Jessenberger, R., A. Dros. Res. Conf. 44:323B.


Michele Vendruscolo, Associate Editor

Accepted September 29, 2003