Surfeit Locus Gene Homologs Are Widely Distributed in Invertebrate Genomes

NIA LL ARMES AND MIKE FRIED*

Eukaryotic Gene Organization and Expression Laboratory, Imperial Cancer Research Fund,
London WC2A 3PX, United Kingdom

Received 27 June 1996/Accepted 22 July 1996

The mouse Surfeit locus contains six sequence-unrelated genes (Surf-1 to -6) arranged in the tightest gene cluster so far described for mammals. The organization and juxtaposition of the genes are conserved between mammals and birds, and this may reflect a functional or regulatory requirement for the gene clustering. We have undertaken an evolutionary study to determine whether the genes are conserved and clustered in invertebrate genomes. Drosophila melanogaster and Caenorhabditis elegans homologs of the mouse Surf-4 gene, which encodes an integral membrane protein associated with the endoplasmic reticulum, have been isolated. The amino acid sequences of the Drosophila and C. elegans homologs are highly conserved in comparison with the mouse Surf-4 protein. In particular, a dilysine motif implicated in endoplasmic reticulum localization of the mouse protein is conserved in the invertebrate homologs. We show that the Drosophila Surf-4 gene, which is transcribed from a TATA-less promoter, is not closely associated with other Drosophila Surfeit gene homologs but rather is located upstream from sequences encoding a homolog of a yeast seryl-tRNA synthetase protein. There are at least two closely linked Surf-3/rpL7a genes or highly polymorphic alleles of a single Surf-3/rpL7a gene in the C. elegans genome. The chromosomal locations of the C. elegans Surf-1, Surf-3/rpL7a, and Surf-4 genes have been determined. In D. melanogaster the Surf-3/rpL7a, Surf-4, and Surf-5 gene homologs and in C. elegans the Surf-1, Surf-3/rpL7a, Surf-4, and Surf-5 gene homologs are located on completely different chromosomes, suggesting that any requirement for the tight clustering of the genes in the Surfeit locus is restricted to vertebrate lineages.

The mouse Surfeit locus contains an unusually tight cluster of six sequence-unrelated housekeeping genes (Surf-1 to -6) and encompasses approximately 45 kb of genomic DNA (5, 9, 15). The relatively high gene density within the Surfeit locus (an average of one gene per 7.5 kb) compared with the whole mouse genome, the alternation of the direction of transcription of neighboring genes (5, 9), the presence of a bidirectional promoter between the 5′ ends of the Surf-1 and Surf-2 genes (13), the overlap of transcripts of the Surf-2 and Surf-4 genes (20), and the conservation of the organization of the locus in higher vertebrates (3, 22) have prompted the suggestion that stringent selection pressures may exist to hold the genes together for regulatory or functional reasons (3, 9).

The Surf-3 gene encodes the L7a ribosomal protein (rpL7a) (6). Biochemical analyses have so far failed to identify the functions of the other Surfeit genes. In the absence of this functional information we have sought other ways in which to assess the importance of the structure and gene organization of the Surfeit locus. One possible way to assess a requirement for specific gene arrangements may be to define the extent to which a given arrangement has been conserved through evolution. The Surfeit locus structure is conserved between mammals (22) and chickens (3). If regulation does exist between the Surfeit genes and the origin of the Surfeit locus predates the separation of invertebrate and vertebrate lineages, then it is feasible that aspects of the locus might be retained in invertebrates.

We have set about identifying and characterizing Surfeit gene homologs in the invertebrates Drosophila melanogaster and Caenorhabditis elegans, not only to evaluate the conservation of the Surfeit locus gene organization, but also to provide a genetic system to aid functional characterization of the Surfeit gene products. We have previously shown that the Drosophila homolog of the Surf-3/rpL7a gene is not closely associated with other Surfeit gene homologs (1). However, we do not know whether other members of the Surfeit locus are clustered in the Drosophila genome or whether the organization of the Surfeit locus is conserved in C. elegans. As it has been estimated that 25% of all C. elegans genes are arranged in cotranscribed gene clusters, some containing functionally related genes (23), it is of considerable interest to determine whether or not Surfeit gene homologs are closely linked in the C. elegans genome.

The mammalian Surf-4 gene encodes a polytopic integral membrane protein containing a dilysine motif near to its C terminus and is located in the endoplasmic reticulum (ER) (18). This dilysine motif is implicated in conferring ER localization to the protein (16). In this work we show that the amino acid sequences of the Drosophila and C. elegans Surf-4 gene homologs, including the dilysine motifs, are highly conserved. The Drosophila Surf-4 gene homolog was cloned, its chromosomal location was analyzed, and the gene was found not to be closely associated with other Drosophila Surfeit gene homologs. We have also determined the chromosomal locations of the C. elegans Surf-1, Surf-3/rpL7a, and Surf-4 gene homologs, and we show that these genes and the C. elegans Surf-5 gene are located at different chromosomal positions and not linked as in higher vertebrates.

MATERIALS AND METHODS

PCRs and oligonucleotides used. A 600-bp biotinylated fragment corresponding to part of a D. melanogaster Surf-4 cDNA clone was amplified from a 12-
24-h developmental stage cDNA library under the following conditions. Reaction mixtures contained the degenerate oligonucleotides 5‘D4GENI (biotin-GA GGAGGIAAGTTGACATCTCCAGTGSG) and 5‘R4AN (CCICATTACG ICATGTGCTGAAATACRATTTT) at 1 μM, each of the four deoxy-nucleoside triphosphates (dNTPs) at 2 mM, KCl at 50 mM, CaCl2 at 1 mM, and 5 U of AmpliTaq (Perkin-Elmer Cetus) and 10 ng of amplified cDNA library. PCR was carried out with a denaturing temperature of 94°C, an annealing temperature of 45°C, and an extension temperature of 72°C, all steps performed for 10 s in a Biometra Trio thermoblock. After 30 cycles products were separated on a 1% agarose gel. The 600-bp product was then gel purified, reamplified for 10 cycles, and sequenced directly by using the protocol supplied with Dynabeads from Dynal.

The probe corresponding to the open reading frame encoding the serl-1 RNA was amplified by PCR from genomic libraries DROSFIN1 (GAATGCGGCGACATTCTG) and DROSFIN2 (AGCGGAGCG CCGTTATC). The C. elegans Surf-3p (pL7a) probe was amplified from C. elegans genomic DNA with oligonucleotides C6G21 (GGCGCGAATCTCCGCTCCA CTTGTTACGC) and C6G22 (GCGCGGCAACCCATACCTGCTCCCCAGT G), which were designed from the sequence with GenBank accession number M79759 and contain additional EcoRI and HindIII sites, respectively, for use in subsequent cloning.

Hybridizations with libraries and DNA and RNA blots. Southern blotting, Northern (RNA) blotting, and slot blotting were performed by using the standard protocols suggested in the Hybond-N protocol booklet from Amersham, Northern (RNA) blotting, and slot blotting were performed by using the standard protocols suggested in the Hybond-N protocol booklet from Amersham, and Hybrid-N membrane was used in all cases. A C. elegans YAC blot was a kind gift of Alan Coulson. DNA probes were labeled by random priming and hybridized with membranes under standard conditions. DNA hybridizations were usually performed at 65°C, and plots were washed in 0.1 × (1 × SSC) at 0.15 M NaCl plus 0.1 M sodium citrate–0.1% sodium dodecyl sulfate (SDS). 5′SSC at 37°C, except for the low-stringency hybridization using the serl-1 RNA probe. The synthetic-encoding open reading frame probe and hybridization of the C. elegans YAC filter with the mouse Surf-1 and Surf-2 probes, which were performed at 55°C and for which plots were washed in 0.05 × SSC at 58°C. Riboprobes used in Northern blotting were amplified from the Eco473D clone using the T7 promoter present in the vector pBR420. Hybridizations were performed as for Southern blots except that 50% formamide was included. Washes were performed in 0.1 × SSC at 65°C.

(continued)
stringency PCRs on a Drosophila 12- to 24-h developmental stage cDNA library in the vector pNB40 (see Materials and Methods). A 600-bp PCR product was produced and sequenced directly. The sequence showed significant homology to those of the mammalian and C. elegans Surf-4 homologs. The PCR product was cloned, labelled, and used to screen 80,000 colonies of a 12- to 24-h developmental stage D. melanogaster cDNA library. Eight positive clones were identified, of which three were characterized further. All three positive DNAs gave common-sized fragments on digestion with Sau3A restriction enzyme. The clone containing the largest insert was sequenced in its entirety and designated DS4cDNA. The DS4cDNA clone possesses a long open reading frame which specifies a protein highly homologous to other Surf-4 homologs (Fig. 1). The D. melanogaster Surf-4 homolog is 59% identical and 79% similar to the murine Surf-4 gene.

The DS4cDNA cDNA clone was labelled and hybridized with a cosmid library of D. melanogaster genomic DNA with a coverage of four genomes (8). Four independent positive cosmids were recovered from secondary screens. The cosmids designated 98F10 was investigated most thoroughly for convenient restriction fragments on Southern blots which hybridized with a DS4cDNA probe. A strongly hybridizing 4.0-kb SacI fragment was subcloned and sequenced and found to contain the entire Drosophila Surf-4 gene (Fig. 2). The Drosophila gene contains three introns (Fig. 2), the first being the largest, with a size of 755 bp, and located in the same relative position as the first introns of the mouse (9) and C. elegans (see above) Surf-4 genes. The second intron (65 bp) is in the same position in the gene as the third intron in the mouse gene. The last intron (62 bp) is in a novel position in respect to the locations of the C. elegans (see above) and mouse introns.

Southern blot analysis was performed on genomic DNA extracted from Drosophila SL2 cells after digestion with three restriction enzymes using DS4cDNA as a probe. Single hybridizing bands were detected for two of the digests, i.e., those with SacI and SpeI, and two bands were detected for a third, with ClaI, known to cut within the Drosophila Surf-4 gene (data not shown). These results indicate that the Drosophila Surf-4 gene is present as a single copy in the fly genome.

**Transcription start point determination for the D. melanogaster Surf-4 gene.** The transcription start sites of the D. melanogaster Surf-4 gene were determined by S1 analysis carried out on poly(A)-enriched RNA from SL2 cells. Single-stranded S1 probes were generated as detailed in Materials and Methods. After digestion with S1 nuclease, the protected fragments were separated on a 6% denaturing polyacrylamide gel alongside a sequenced M13 template used as a size marker. Although transcription start sites were detected over a region of about 54 bp (nucleotides 645 to 698 in Fig. 2), one small region (nucleotides 675 to 677 in Fig. 2) accounted for a large proportion of start sites and may be considered to contain the major transcriptional start sites. Transcription start sites determined by 5’ RACE (1) confirmed the S1 analysis (data not shown). Sequence analysis of this region indicates that there are no TATA boxes positioned within suitable range of these start sites (Fig. 2).

**Analysis of the genomic DNA surrounding the D. melanogaster Surf-4 gene.** The 4.0-kb genomic SacI fragment contains 700 bp downstream from the Surf-4 gene polyadenylation site. A further 2.1 kb of DNA downstream of the 4.0-kb SacI fragment was subcloned and sequenced. No sequences within this 2.8 kb of genomic sequence showed homology to any of the known mouse Surf4.5 genes. However, when used to search the DNA and protein databases, these sequences detected significant homology to a Saccharomyces cerevisiae seryl-tRNA synthetase gene. Figure 3 shows an alignment of the conceptual protein encoded by sequences downstream of the Drosophila Surf-4 gene with the S. cerevisiae seryl-tRNA synthetase protein. The conceptual translation product shown in Fig. 3 is derived from part of a long open reading frame in the 2.8 kb of sequence downstream of the Surf-4 gene stretching from the distal EcoRI site and extending 1,236 bp towards the 3′ end of the Surf-4 gene (Fig. 2). Most of the amino acid sequence of this long open reading frame had significant linear homology with the S. cerevisiae seryl-tRNA synthetase protein, although no homology was detected among approximately the first 100 amino acids of the open reading frame extending from the EcoRI cloning site. The Drosophila translation product possesses 17 more amino acids at its C terminus than does the yeast homolog. A PCR product encompassing most of the DNA homologous to seryl-tRNA synthetase was generated, labelled, and hybridized with a Southern blot of D. melanogaster genomic DNA. The probe detects only single bands in genomic DNA digested with SacI, ClaI, and SpeI restriction enzymes (data not shown). This indicates that the seryl-tRNA synthetase gene is present as a single copy in the fly genome.

**Chromosomal location of the D. melanogaster Surf-4 gene.** The 4.0-kb SacI fragment containing the entire Surf-4 gene was used as a probe for in situ hybridization with salivary gland polytene chromosomes. The probe hybridized with band 88F on the third chromosome around subband 3 or 4. There are several genes required for the correct development of the D. melanogaster flight muscles in this region which have been well characterized. The entire 88F region has been cloned in a phage lambda clone contig (10). Samples of these lambda clones were obtained and used to determine the precise location of the D. melanogaster Surf-4 gene. A selection of lambda clones were slot blotted onto Hybond-N, and the DS4cDNA clone was labelled and hybridized with this blot (Fig. 4). Two of the lambda clones gave positive signals (TM325 a strong signal and TM16 a much weaker signal), allowing precise localization of the Drosophila Surf-4 gene within the lambda contig. Flybase was consulted in order to determine the existence of known genes, mutants, or other aberrations in this region. Several P-element-induced lethal mutants in the vicinity of the Surf-4 gene were obtained and were investigated by Southern blotting for insertion near to the Surf-4 gene (see Materials and Methods). No insertions were detected in the direct vicinity of the Surf-4 gene.

C. elegans contains at least two Surf-3/rpL7a genes. C. elegans homologs for Surf-4 (see above) and Surf-5 (5) have been identified. We therefore attempted to identify and localize a C. elegans Surf-3/rpL7a homolog. Two oligonucleotides specific for the C. elegans Surf-3/rpL7a gene were designed by using sequences present in the DNA databases generated from expressed sequence tags (see Materials and Methods). These oligonucleotides were used to amplify an approximately 300-bp fragment(s) from C. elegans genomic DNA. The fragments were cloned and two clones were sequenced. The two sequences were capable of encoding the expected portion of the Surf-3/rpL7a protein but contained an additional internal sequence of 52 bp in one clone and 54 bp in the other (Fig. 5). These additional sequences were present in the same relative location in both clones and most likely represent introns, as they are flanked by consensus splice donor and acceptor sequences and are located at the same position as is intron 6 of the mouse, human, and chicken Surf-3/rpL7a genes (2, 6, 14). However, these intervening sequences differed not only by 2 bp in length but also by three mismatches. Seven other sequence mismatches were found between the two C. elegans Surf-3/rpL7a sequences in the coding regions, of which only one...
change alters the encoded amino acid. Most of these differences seem unlikely to be generated during PCR, as one of the products was identical in the coding region to the sequence tag from which the oligonucleotides were derived and because deletion-insertions are rare PCR errors. These results suggest that there are at least two Surf-3/rpL7a genes in the C. elegans genome or that highly polymorphic alleles of this gene are present.

Three C. elegans Surf gene homologs are not linked in the C. elegans genome. One of the C. elegans Surf-3/rpL7a PCR products was hybridized with a gridded YAC filter containing inserts covering the whole C. elegans genome (see Materials and Methods). Two overlapping YACs (Y48F1 and Y60B1) were detected by the probe, and consultation of the C. elegans database indicated that the gene(s) was located on chromosome IV. The fact that the two YACs detected by the probe are overlapping suggests that if there is more than one copy of this gene present, then the copies are closely linked and not dispersed in the genome.

No positive YACs were detected by using mouse Surf-1 and Surf-2 cDNAs to probe the C. elegans YAC filter under low-stringency conditions. A search of the C. elegans database with the mouse Surf-1 cDNA sequences detected a mouse Surf-1 homolog on cosmid KK5K1, which maps to C. elegans chromosome V. No sequences with sequence homology to the...
The Surf-4 gene has been mapped to the X chromosome. In mammals and chickens the Surf-5 gene lies 5' to the Surf-3/rpl7a gene (3, 5, 9, 21, 22). No sequence homology to the mouse Surf-3/rpl7a gene was found in cosmid CEZK970. In summary, four Surfet gene homologs in C. elegans have been identified. These four genes are not linked in the C. elegans genome, being located on completely different chromosomes, with the Surf-5 gene on chromosome II, the Surf-1 gene on chromosome V, the Surf-3/rpl7a gene on chromosome IV, and the Surf-4 gene on chromosome X.

**DISCUSSION**

The Surfet locus contains the tightest cluster of mammalian genes so far described. The organization and juxtaposition of at least five of the Surfet genes are conserved between mammals and birds (600 million years of divergent evolution), possibly as a result of selective pressures which have prevented disruption of the locus structure. We are investigating the possibility that the unusual organization of the Surfet genes and the conservation of the locus structure underlie deeper regulatory or functional relationships between the individual genes. We have here analyzed two metazoans that are phylogenetically more distant from mammals than birds, namely, D. melanogaster and C. elegans, for the presence of Surfet gene homologs and also to determine whether identified homologs are physically associated. This information should aid consideration of the evolutionary significance of the locus structure and help to address the question of when the locus structure arose. We have previously analyzed the genomic DNA surrounding the Drosophila homolog of the Surf-3/rpl7a gene and shown that it is not associated with other Surfet gene homologs. In the present work we have isolated highly conserved Drosophila and C. elegans homologs of the mammalian Surf-4 gene and shown that they too are not associated with other invertebrate Surfet gene homologs.

The mouse Surf-4 gene encodes an integral membrane protein associated with the ER. The human and mouse Surf-4 proteins possess a dilyssine motif at their C termini, a motif which has been demonstrated to impose ER localization on heterologous proteins and is present in many ER-resident proteins (16). The Drosophila and C. elegans Surf-4 homologs possess C-terminal dilyssine motifs at locations equivalent to those of the mammalian proteins, and preliminary experiments involving transfection of an expression construct containing an epitope-tagged Drosophila Surf-4 gene suggest that the Drosophila Surf-4 protein is also resident in the ER (data not shown). This suggests that the Surf-4 protein has a conserved cellular location and supports the possibility that this dilyssine motif may serve as an ER targeting signal in all metazoan Surf-4 homologs.

We have mapped the Drosophila Surf-4 gene to the polytene band 88F, a region that has previously been entirely cloned. The precise location of the Surf-4 gene was refined by analysis of cloned DNA fragments. However, no previously characterized mutations mapping to this region result from disruption of the Surf-4 gene. This result has prevented us from undertaking further genetic analysis of this gene in flies at this time.

In D. melanogaster the Surf-4 gene is a single-copy gene, as is the case in mice. However, whereas in mice the 3' ends of the Surf-4 and Surf-2 genes overlap, analysis of the genomic region downstream of the Drosophila Surf-4 gene has revealed that it is not closely associated with a fly Surf-2 homolog, or indeed with any Surfet gene homolog. Instead, an open reading frame which could encode a homolog of the yeast seryl-tRNA synthetase is located directly downstream of the Drosophila gene.
Two other Surfeit gene homologs have been detected in *D. melanogaster*, and these are located at chromosomal locations different from that of the Surf-4 gene. The *Drosophila* Surf-3 gene is found at band 6 A/B on the X chromosome (1), and a *Drosophila* cDNA sequence encoding a Surf-5 homolog has been placed in the DNA database (accession no. L46826) and maps to band 35 A1 on chromosome 2L.

It was hoped that some support for a functional relationship between the Surfeit genes underlying their clustering in higher vertebrates might be found in *C. elegans*, as about 25% of all *C. elegans* genes are closely linked in cotranscribed gene clusters in which the genes may be functionally related (23). We have identified homologs of the Surf-1, Surf-3, Surf-4, and Surf-5 genes in the *C. elegans* genome and find that they are not closely associated, being located on completely different chromosomes. The *C. elegans* Surf-2 gene was not detected but may not be well conserved. Whereas there is high-level amino acid homology between the mouse and human homologs of Surf-3/rpL7a (99.6%) (2), Surf-4 (98.9%) (18), Surf-5 (98.5%) (5), and Surf-1 (79.1%) (12), the Surf-2 homologs of mice and humans are only 68.4% identical at the amino acid level (12). There is some evidence that there may be more than one copy of the Surf-3 gene present in *C. elegans* and that the duplicate copies are closely linked in the genome.

The results presented here clearly demonstrate that the Surfeit genes identified are not tightly clustered in invertebrate organisms, as they are in higher vertebrates. Further work must be undertaken in order to determine whether an ancestral Surfeit gene cluster was disrupted in these invertebrate lineages or whether the Surfeit locus was assembled immediately prior to or during early vertebrate evolution.

This study was performed in order to assess if a close association between Surfeit gene homologs exists in invertebrates. One can reason that a requirement for a specific gene organization will affect the ease with which the genes involved can be dissociated. The HOM-C complex of *D. melanogaster* and the vertebrate Hox loci contain genes which are arranged in a specific conserved order (7), and in this case it is suggested that a requirement for the structures of these loci has enforced their conservation (11). Our failure to detect linkage of Surfeit gene homologs in invertebrates may be evidence against a strong selection pressure maintaining the structure of the Surfeit locus. Alternatively, it is possible that the Surfeit locus structure arose after the divergence of the vertebrate and invertebrate lineages studied here and that only then did a regulatory interrelationship develop between the genes. We cannot yet distinguish between these alternatives. On the other hand, recent progress in mammalian genome mapping suggests that local clustering of unrelated genes may be common in mammalian genomes. Gene-dense regions might arise in vertebrate genomes because the genome expands (or contracts) in a nonrandom fashion and/or because karyotypic rearrangement events favor gene association and clustering. It is now known that housekeeping genes, and those genes located in R bands in the mammalian genome, are generally found in more gene-dense regions and are smaller (i.e., less expanded) (4). This suggests that levels of DNA turnover are not equal throughout the mammalian genome and raises the possibility that the Surfeit locus genes may have associated by the action of genome evolutionary forces of a less specific nature. Our ongoing studies should help to determine whether the Surfeit locus structure is conserved in all vertebrate lineages, address the question of the point at which the locus came into existence, determine any biological significance of the tight clustering, and determine whether the Surfeit genes are functionally related.

ACKNOWLEDGMENTS

We thank Sheena Pinchin for performing the *Drosophila* chromosome mapping; Clive Dickson, David Ish-Horowicz, and Jon Gilley for their useful comment in the preparation of the manuscript; and Alan Coulson for the *C. elegans* YAC filter.

REFERENCES