Intron–exon organization of the gene for the multifunctional animal fatty acid synthase

(discrim boundaries/evolution/gene fusion/domain shuffling/intron phasing)

Christopher M. Amy, Brenda Williams-Ahlf, Jürgen Naggert*, and Stuart Smith†
Children’s Hospital Oakland Research Institute, 747 Fifty-second Street, Oakland, CA 94609

Communicated by Paul K. Stumpf, November 12, 1991 (received for review August 23, 1991)

ABSTRACT The complete intron–exon organization of the gene encoding a multifunctional mammalian fatty acid synthase has been elucidated, and specific exons have been assigned to coding sequences for the component domains of the protein. The rat gene is interrupted by 42 introns and the coding region of the gene, 23 split the reading frame in phase I, 14 split the reading frame in phase 0, and only 4 split the reading frame in phase II. Remarkably, 46% of the introns interrupt codons for glycine. With only one exception, boundaries between the constituent enzymes of the multifunctional polypeptide coincide with the location of introns in the gene. The significance of the predominance of phase I introns, the almost uniformly short length of the 42 introns and the overall small size of the gene, is discussed in relation to the evolution of multifunctional proteins.

The de novo synthesis of fatty acids from malonyl-CoA requires seven enzyme activities that in most bacteria and plants exist as discrete monofunctional polypeptides (1). However, in fungi these activities are distributed between two nonidentical polypeptides encoded in two unlinked genes (2–5), and in animals they are integrated into a single polypeptide chain (6, 7). The hypothesis that the multifunctional fatty acid synthase evolved by fusion of the genes for ancestral monofunctional enzymes is supported by the finding that several of its components can be proteolytically cleaved from the complex as discrete catalytically competent monofunctional proteins that exhibit amino acid sequence similarity with present-day monofunctional analogs (7–10). It has been anticipated that structural features of the animal fatty acid synthase gene such as the location of introns may provide clues as to how the gene fusion events occurred (9). The objective of this study therefore was to determine the nucleotide sequence of the rat fatty acid synthase gene and map the location of the introns, particularly in relation to the location of the various catalytic domains of the multifunctional protein.†

EXPERIMENTAL PROCEDURES

Genomic Clone Isolation and Sequencing. Genomic libraries, prepared from partial EcoRI or Hae III digests of Sprague–Dawley rat DNA cloned into the EcoRI site of the λ vector Charon 4a (11), were screened with probes containing parts of rat fatty acid synthase cDNA clones obtained from a lactating rat mammary gland library (12). Four overlapping clones from the two rat genomic libraries were identified that spanned the entire gene (13). Double-stranded genomic DNAs recloned in pUC vectors were sequenced directly with the T7 sequencing kit (Pharmacia) and adenosine 5’-[α-32P]thiotriphosphate (3000 Ci/mmol; 1 Ci = 37 GBq; Amersham) according to the manufacturer’s instructions.

Location of Exon–Intron Boundaries. All of the introns were sequenced and their boundaries were established by comparison with the cDNA sequence (12).

RESULTS AND DISCUSSION

The rat fatty acid synthase gene was sequenced in its entirety and the location of introns was established (Fig. 1). The coding sequence is interrupted by 42 introns and the sequences bordering the splice-site junctions universally follow the GT/AG rule. However, of the 41 introns that interrupt the coding region of the gene, 23 split the reading frame in phase I, 14 split the reading frame in phase 0, and only 4 split the reading frame in phase II. Remarkably, 46% of the introns interrupt codons for glycine.

It has been pointed out that many mosaic proteins, such as extracellular matrix proteins, blood coagulation proteins, and membrane-associated proteins are composed of modules that appear to have fused in various combinations (15). These modules, which encode discrete functional units such as epidermal growth factor domains, fibronectin domains, and kringle domains, are composed of exons flanked at both ends by introns in the same phase, and it has been suggested that this identical phasing could have promoted the coupling of functional domains through the process of intronic recombination by preserving a continuous open reading frame (15). Although from the sequence data available at present it is not possible to say whether the exons that constitute the multifunctional fatty acid synthase gene represent modules that have been shuffled in various combinations to produce other unique mosaic proteins, the exons flanked by phase I introns clearly cluster together in continuous groups, exons 3–6, 17–21, 26–27, 30–31, and 37–39 (Figs. 1 and 2). Similarly, exons flanked by phase 0 introns cluster together at exons 13–14, 23–24, and 41–42. These observations are consistent with common phasing having played a role in the fusing of some of the domains through intronic recombination.

As is typical for a vertebrate gene (17) the first intron, which separates the single 5’-noncoding exon from the first coding exon, and the last exon are relatively long. Indeed, the entire 3’-noncoding region of the mRNA is represented in a single exon, the largest in the entire gene, showing that the 8.3- and 9.1-kilobase mRNA species transcribed in the rat arise not from alternative splicing but from differential utilization of two polyadenylation signals (12). This conclusion has also been confirmed by amplification of rat spleen DNA by using primers flanking the polyadenylation signal for the

Abbreviation: nt, nucleotides.
*Present address: The Jackson Laboratory, Bar Harbor, ME 04609.
†To whom reprint requests should be addressed.
‡The sequence reported in this paper has been deposited in the GenBank data base (accession no. M84761).
interruption and by phase 0, interruption with the
scripts. Invariant otides (nt) 13154-13254],
smaller mRNA species; only fragments with the lengths
predicted by the cDNA and genomic clone sequences were
produced (details not shown). Two elements of the short
interspersed repetitive DNA sequence type (16, 18) are
associated with the fatty acid synthase gene, one between
the enoyl reductase and ketoreductase coding sequences [nucleotides (nt) 13154-13254], the other, in the opposite orienta-
tion, in the 5' flanking region (nt -3724 to -3833). The
elements exhibit 91% sequence identity, both contain A-rich 3' ends, and both are flanked by direct DNA repeats (AAACCTGATGACCC and AGA>CATT, respectively).

Earlier we proposed precise locations for the functional
components of the animal fatty acid synthase based on the
hypothesis that the individual catalytic domains should ex-
hibit significant sequence similarity with monofunctional
analogous and might be expected to be joined by mutable
surface loops, which would be susceptible to proteolytic
attack (10). Indeed, similar features have been found at
the boundaries of other multifunctional proteins (19) and it seems
likely that fusion of the genes for the primordial monofunc-
tional proteins in a way that results in joining of the poly-
peptides via a hydrophilic surface loop region is likely to
preserve the three-dimensional structure and biological func-
tion of each of the fused components. Of the seven functional
components of the animal fatty acid synthase, only the
dehydrase could not be precisely located, although we sug-
gested that it was likely to be positioned somewhere between
the transferase and enoyl reductase domains (10). More
recently, S. Donadio and L. Katz (personal communication)
and P. F. Leadlay and colleagues (personal communication)
have attempted to locate the interdomain boundaries of
the multifunctional polyketide synthase of *Saccharopolyspora
erythraea* by similar criteria. The predicted amino acid se-
quence, the size, and the organization of the individual
domains of this protein are strikingly similar to that which we
have elucidated for the animal fatty acid synthase, except
that the polyketide synthase likely contains not one but
multiple fatty acid synthase-like modules, and several of
these modules appear to lack specific functional domains.
The conclusions reached by these workers concerning the
depicted size and location of the functional domains of the
polyketide synthase agree remarkably well with the model we
advanced earlier for the animal fatty acid synthase. In
addition, since in the case of the polyketide synthase only one
of the modules contains a dehydrogenase function, S. Donadio
and L. Katz (personal communication) were able to deduce
that this enzymatic function is encoded within 140-170 amino
acids immediately downstream of the transferase domain. A
stretch of 100 amino acids within the corresponding region of
the animal fatty acid synthase (residues 875-975) exhibits
marginally significant sequence similarity with both the pu-
tative dehydrogenase region of the polyketide synthase and the
170-residue monofunctional β-hydroxydecanoyl thioester de-
hydrolase that provides the branch point to unsaturated fatty
acids in *Escherichia coli* (ALIGN scores 3-4 SD). This evi-
dence supports our original suggestion that the dehydro-
dray domain of the animal fatty acid synthase is located between
the transferase and enoyl reductase and suggests that the
most likely location is to the C-terminal end of the trans-
ferase in a region encoded by exons 16-19. Again, the
boundaries of the encoded putative dehydrogenase domain are
marked by a predicted surface region at the N-terminal end
and by a highly mutable predicted surface region at the C-
terminal end (details not shown).

Because of the tentative nature of the evidence relating to
the location of the dehydrogenase domain, it is not possible to
pinpoint precisely the location of the associated boundaries.
However, for the remaining functional components of the
complex their boundaries can be predicted with confidence
and, with only one exception, appear to coincide with the
location of introns in the gene (Fig. 2). The 5' ends of exons
28 and 37 and the 5' ends of exons 33 and 39 encode plausible
amino acid sequences, which, we have deduced, define
interdomain boundaries (10). Thus, exons 29-32 appear
to encode the enoyl reductase, exons 33-37 encode the ketore-
ductase, exon 38 encodes the acyl carrier protein, and exons
39-42 encode the thioesterase; exon 42 also includes the

---

**Fig. 1.** Location and characteristics of introns in the rat fatty acid
synthase gene. The position of introns between codons is indicated
by phase 0, interruption after the first base is indicated by phase I;
and interruption after the second base is indicated by phase II; the
amino acid(s) encoded at the splice site is indicated. The locations
of the splice junctions in the cDNA sequence are indicated as super-
scripts. Invariant nucleotides at the splice sites are in boldface type.
FIG. 2. Map of the rat fatty acid synthase gene and the encoded multifunctional polypeptide. Exons are shown as boxes separated by introns depicted as lines. Exons flanked at both 5' and 3' ends by phase 1 introns are solid boxes, those flanked at both ends by phase 0 introns are darkly shaded boxes, and those flanked by introns of mixed phasing are lightly shaded boxes. SINE, short interspersed repetitive DNA sequence (16). The location of interdomain boundaries in the multifunctional polypeptide (10) is shown in relation to the position of introns in the gene. The locations of active site residues for the ketoacyl synthase, transferase, and thioesterase; the nucleotide-binding site motifs; and the site of 4'-phosphopantetheine attachment in the acyl carrier protein (ACP) domain are also marked; the location of the dehydrase coding region is tentative (see text). Exons, introns, and protein domains are drawn to scale. kb, Kilobases.

entire 3'-noncoding region. A mutable amino acid sequence corresponding to the C-terminal end of the malonyl/acytetyl transferase is also encoded in a region of the gene interrupted by an intron (intron 15). However, there is no intron located in the region of the gene encoding the mutable amino acid sequence marking the boundary between the transferase and the ketoacyl synthase; this boundary is apparently encoded =200 nt from the 5' end of exon 9. Exon 9 is, however, the largest in the coding sequence and it is possible that an intron marking this interdomain boundary has been lost. The central region of the fatty acid synthase, encoded by exons 20–28, has not yet been assigned a functional role. The region is the least conserved in the multifunctional polypeptide and may play a role in stabilizing subunit interactions (10). An extensive, poorly conserved region of unknown function is also present in the corresponding location of each of the fatty acid synthase-like modules that constitute the polyketide synthase of S. erythraea (20).

The two readily identifiable reductases contain, in their first exons, the Gly-Xaa-Ser-Xaa-Gly motif characteristic of nucleotide-binding domains (10). The coding sequences for the motif in the enoyl reductase and ketoreductase domains begin 108 nt into exon 29 and 109 nt into exon 33, respectively. The two nucleotide-binding domains exhibit significant sequence homology with the corresponding domains of other pyridine nucleotide-requiring enzymes, and secondary structure analysis predicts the presence of a βαβ subdomains typical of a mononucleotide-binding fold (10). In common with other pyridine nucleotide enzymes, the nucleotide-binding structural modules of the multifunctional fatty acid synthase are not encoded by single exons. However, the sizes of the introns within the reductase domains of the fatty acid synthase are considerably smaller than those found, for example, in other vertebrate dehydrogenases (21). If the nucleotide-binding domains of the multifunctional fatty acid synthase are descendants of the same primitive mononucleotide-binding fold, encoded within several exons, that has fused with a variety of substrate-binding domains to give rise to many different enzymes with different specificities (22), then it would appear that the introns within the reductase coding regions must subsequently have been shortened. Indeed, perhaps the most remarkable feature of the fatty acid synthase gene is the overall relatively short length of the introns. The mean length of introns that interrupt the coding regions of vertebrate genes has been estimated as 1127 nt (17); in contrast, the mean length of introns within the coding region of the fatty acid synthase gene is only 191 nt. The significance of the prevalence of shortened introns within the gene is not immediately clear. On fusion of the seven genes encoding the component activities of the fatty acid synthase it was probably advantageous that individual promoter and enhancer regions be removed to avoid deleterious effects resulting from the presence of multiple regulatory sequences positioned throughout the gene. This process could have resulted in a shortening of a number of the intronic regions. However, the fatty acid synthase introns are almost uniformly short, whether they are located at the 5' end of individual enzyme coding regions or whether they interrupt the internal coding regions of the enzymes, suggesting that perhaps there has been additional selective pressure, either on the size of individual introns or on the upper size limit for the gene. Other possible pressures might have been related to the chromosomal location of the gene or the need to have a compact gene that could be transcribed quickly and efficiently. Packaging of a highly fragmented gene, which encodes a long mRNA into a relatively modest stretch of DNA could avoid the extensive chromatin decondensation that would otherwise be required for activation of a gigantic gene and might allow more rapid transcription in response to regulatory stimuli. The fatty acid synthase gene is highly regulated in a tissue-specific manner according to the nutritional, hormonal, and developmental status of the animal, and both its mRNA and the encoded protein are highly abundant in tissues active in lipogenesis (13).

Fatty acid synthases have been found in eukaryotes in two multifunctional forms, the α2β2 yeast form. The component enzymes of the two types of complex exhibit little sequence similarity and are connected structurally in completely different ways [in yeast: α, acyl carrier protein–ketoreductase–ketoacyl synthase; β, acetyl transferase–enoyl reductase–dehydrase–malonyl/palmitoyl transferase (23)], indicating that they have most likely evolved by separate series of gene fusion events. Until recently, it had seemed likely that the two independent series of gene fusions took place after divergence of the yeast and animal lines. However, the finding that the multifunctional proteins required for polyketide synthesis in S. erythraea and fatty acid synthesis in animals are composed of functional domains of similar size and amino acid sequence linked in the same order [ketoacyl synthase–transferase–dehydrase–enoyl reductase–ketoreductase–acyl carrier protein (20, 24)] directly challenges this view and raises the distinct possibility that the gene fusion events leading to a common polyketide synthase/fatty acid synthase module took place prior to the divergence of prokaryotes and eukaryotes. Most of the clearest examples of exon shuffling involve genes that are unique to vertebrates and that presumably appeared relatively late in eukaryotic evolution (15, 25). However, if the introns located at the
boundaries of the coding regions for the constituent enzymes of the fatty acid synthase are indeed relics of the process of assembly of the multifunctional polypeptide, then it seems likely that this assembly process occurred early in evolution, before separation of the prokaryote and eukaryote lines. Of course, the possibility that the animal-type fatty acid synthase module was acquired by prokaryotes by a relatively late "horizontal" gene transfer mechanism cannot be entirely ruled out at present but could be evaluated once additional sequences of related polyketide and fatty acid synthases become available.

We thank Drs. R. F. Doolittle (University of California, San Diego) and K. Paigen (The Jackson Laboratory, Bar Harbor, ME) for their helpful comments. We also are very grateful to Drs. S. Donadio and L. Katz (Abbott Laboratories, IL) and Dr. P. F. Leadlay (University of Cambridge, Cambridge, U.K.) for generously sharing with us their unpublished results. The work was supported by Grant DK 16073 from the National Institutes of Health.