Relationships between the Gene and Protein Structure in Human Complement Component C9[†,‡]

Daniela Marazziti,[§] Gösta Eggertsen,[‖,⊥] Georg H. Fey,[‖] and Keith K. Stanley*[,§]

European Molecular Biology Laboratory, Meyerhofstrasse 1, Postfach 10.2209, D-6900 Heidelberg, FRG, and Research Institute of Scripps Clinic, 10666 North Torrey Pines Road, La Jolla, California 92037

Received February 3, 1988; Revised Manuscript Received May 2, 1988

ABSTRACT: Human complement component C9 is a multidomain protein for which a large number of surface topographical features have been determined. We have analyzed the exon–intron boundaries of the human C9 gene and find a good correlation between splice sites and surface features of the protein but little correlation with the putative protein domain structure, even in the cysteine-rich sequence homology with the low-density lipoprotein (LDL) receptor which is likely to be an independently folded structural motif. This is surprising because in the LDL receptor the same sequence is precisely bounded by introns, and it has been assumed that this sequence is present in both proteins as a result of exon shuffling. We deduce that substantial rearrangement of the exon–intron structure of the C9 gene must have occurred before the exchange of cysteine-rich domains, possibly linked to the process of exon duplication which was required to generate the repeats in the LDL receptor.

Complement component C9 is an interesting protein in which to test the possible relationships between exon–intron boundaries and protein structure. Although it has not yet been crystallized, a wealth of structural features has been determined by biochemical (Biesecker et al., 1982) and recombinant DNA experiments (Stanley et al., 1982; Stanley & Herz, 1987). These studies suggest that there are at least five independently folding domains (Stanley & Herz, 1987), consistent with electron micrographs of the protein (Podack & Tschopp, 1982a,b). One of these domains, which is resistant to proteolysis after C9 inserts into the membrane (Hammer et al., 1977) and is labeled by membrane-restricted photoaffinity probes (Schäfer et al., 1987), presumably corresponds to the membrane inserting region of the molecule. Two of the other putative domains are deduced from sequence homologies. These homologies show that C9 is a mosaic protein with short sequence units homologous to repeating sequences found in thrombospondin (Lawler & Hynes, 1986; DiScipio et al., 1987), the LDL receptor (Stanley et al., 1985), and the EGF precursor (Doolittle, 1985; Stanley et al., 1986). In each of these homologous regions six cysteine residues are present in a span of 40–60 amino acids. Because of the repetitive nature of these segments, their presence in unrelated proteins, and comparison with the known structure of multidomain proteins containing cysteine-rich sequences [e.g., wheat germ agglutinin (Drenth et al., 1980)], it has been suggested that these sequences represent independently folding structural motifs of the protein (Stanley et al., 1986). The cysteine-rich sequence with homology to the LDL receptor (called the class A cysteine-rich sequence motif to avoid confusion with the "growth factor" repeat also found in both molecules) is particularly interesting because the intron–exon boundaries align precisely with the start of the putative protein domain (Südho. et al., 1985a). This observation has therefore supported the hypothesis that exons encode independently folding elements of a protein which may be shuffled between genes by recombination events occurring in intron sequences, thus enabling a rapid diversification of protein structure (Gilbert, 1978; Blake, 1979). Indeed, it has been predicted that the cysteine-rich sequences of the C9 gene will be found on similar discrete exons (Südhof et al., 1985b). We have cloned and sequenced the parts of the C9 gene encoding these regions and examine the relationship between exon–intron boundaries and the putative structure of C9.

Experimental Procedures

Genomic clones were isolated from human genomic libraries constructed in EMBL 3 (Frischauf et al., 1983), in pcos 2 EMBL (Poustka et al., 1984), and in Charon 4 (Maniatis et al., 1978). Further library screenings were performed with human genomic libraries constructed in pHC79 (Hohn & Collins, 1980) and in a second library in EMBL 3 which was a gift from H. Haymerle.

C9 cDNA probes were labeled by nick translation (Rigby et al., 1977) or by random primer extension (Feinberg & Vogelstein; 1984) in order to screen the libraries by standard methods (Maniatis et al., 1982). Restriction maps of the isolated genomic clones were constructed by single and double restriction enzyme digests or by the partial digest method of Rackwitz et al. (1984). Southern blot hybridization (Southern, 1975; Church & Gilbert, 1984) was used to identify fragments containing particular exons.

DNA fragments containing exon sequences were excised from λ or cosmid clones and randomly subcloned into mp8, mp9 or mp18, mp19 (Norrander et al., 1983). Single-stranded DNA of M13 recombinant phages was prepared as described (Sanger et al., 1980), and exon-containing clones were identified by hybridizing a full-length cDNA probe to 5 μL of phage supernatant spotted on nitrocellulose filters. Positive

[†] This work was supported by a fellowship from the Deutscher Akademischer Austauschdienst (to D.M.), Grant R01 AI19651 from the National Institute of Allergy and Infectious Diseases (to G.H.F.), and a stipend from the Swedish Physicians Society (to G.E.).

[‡] The nucleic acid sequence in this paper has been submitted to GenBank under Accession Number J02833.

* Author to whom correspondence should be addressed.

[§] European Molecular Biology Laboratory.

[‖] Research Institute of Scripps Clinic.

[⊥] Present address: Department of Clinical Chemistry I, Huddinge University Hospital, 14186 Huddinge, Sweden.

0006-2960/88/0427-6529$01.50/0

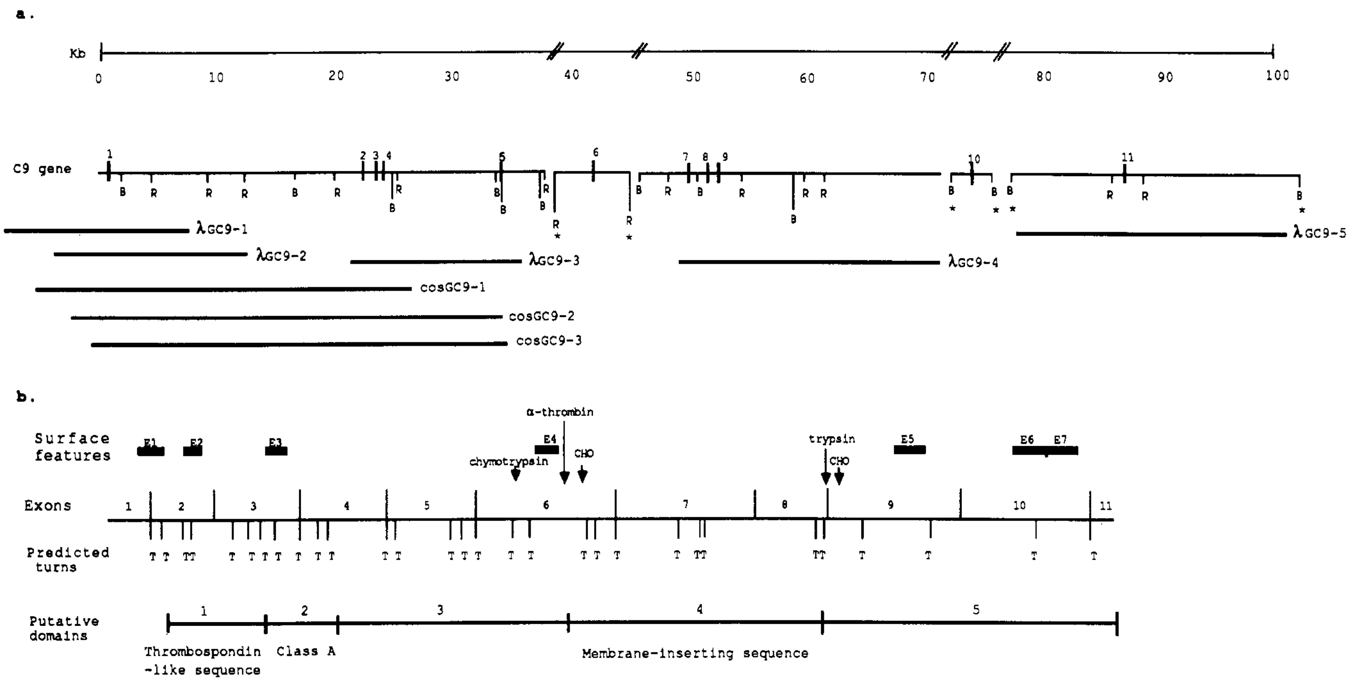


FIGURE 1: (a) Maps of the genomic clones and structure of the C9 gene. Relative positions of λ and cosmid genomic clones are shown in a linear map of the gene. Restriction endonuclease sites *Eco*RI and *Bam*HI are marked R and B, respectively. Exons are numbered 1–11, although it was not excluded that exons 6 and 10 could be further divided by intron sequences. Where fragment sizes were determined from genomic Southern blots, the sites lie outside the λ and cosmid clones and are marked with an asterisk since the precise position is unknown. (b) Correlation of exon–intron boundaries with protein structure. The sequence positions at which introns occur in the C9 gene are shown relative to surface features of the protein and cysteine-rich sequence homologies. E1 to E7 are epitopes of different anti-C9 antibodies, CHO marks the location of N-linked oligosaccharide, and T shows the position of predicted turns.

clones were sequenced by the method of Sanger et al. (1977).

Results

Isolation and Characterization of C9 Genomic Clones. Two different human genomic λ libraries (Maniatis et al., 1978; Frischauf et al., 1983) containing 1.6×10^6 and 7.5×10^5 clones were screened by in situ DNA hybridization with probes derived from C9 cDNA clones (DiScipio et al., 1984; Stanley et al., 1985). Five positive clones were isolated and mapped to regions of the C9 sequence by Southern blot hybridization, by use of probes from different parts of the cDNA.

λGC9-1 and λGC9-2 were found to be overlapping and to contain sequences at the 5′ end of C9 cDNA; λGC9-3 was a nonoverlapping clone containing the sequence coding for the class A cysteine-rich motif (Figure 1a). λGC9-4 contained sequences encoding the central part of C9 including the membrane inserting domain, while λGC9-5 overlapped with the 3′ end of the coding region (Figure 1a).

A human genomic cosmid library of 2×10^5 colonies (Poustka et al., 1984) was also screened with a full-length cDNA clone. Three positive clones were isolated (cosGC9-1, cosGC9-2, and cosGC9-3 in Figure 1a) which all contained sequences from the 5′ end of C9.

Restriction analysis and Southern blot hybridizations revealed that the λ and cosmid clones were not all overlapping, leaving two uncloned regions of the C9 gene (Figure 1a). Despite several further screening experiments with these and different libraries (see Experimental Procedures), no additional clones could be obtained, suggesting that the remaining regions might be unstable in λ and cosmid vectors (Wyman et al., 1986).

All the available λ and cosmid clones were mapped with respect to *Eco*RI and *Bam*HI restriction endonuclease sites (Figure 1a), and cloned fragments hybridizing with C9 cDNA were then compared with hybridizing bands on genomic Southern blots of human white cell DNA after restriction with the same enzymes. In this way the *Bam*HI and *Eco*RI bands corresponding to the two missing exons 6 and 10 were identified and then distinguished with a probe specific for exon 10. The positions of exons 2, 3, 4, 5, and 11 were determined relative to the DNA nucleotide sequences. Exons 7–10 were only mapped to restriction endonuclease fragments on the map in Figure 1a.

All the bands observed on a genomic Southern blot could be assigned to a different part of the C9 gene suggesting that only one C9 gene is present in the human genome.

Exon–Intron Organization. Fragments of genomic clones which hybridized to the cDNA probes were in some cases subcloned in pBR322 or else directly shotgunned into M13. Clones containing exon–intron boundaries were identified as described under Experimental Procedures. These clones were then sequenced, and the exon–intron boundaries were determined by aligning the genomic and the cDNA sequences. In this way it was shown that the C9 gene contained at least 11 exons. For nine of these the complete exon sequence as well as exon–intron boundaries was determined, thus excluding the possibility of further introns in these regions (Table I).

In all cases the consensus boundary sequence (Mount, 1982) was observed (Table I). Although the exon size was fairly similar (100–250 bp), introns varied from 250 bp to over 20 kb. This gave rise to a striking distribution of exons in the gene with clusters of closely spaced exons separated by large introns. The exon sequences are in agreement with the previously reported cDNA sequence (DiScipio et al., 1984; Stanley et al., 1985), with minor variations. At position 22 of the mature protein a cysteine residue (Stanley et al., 1985) rather than an arginine (DiScipio et al., 1984) was found giving an even number of cysteine residues. In the membrane inserting region an additional valine was found at position 293, consistent with the presence of a hydrophobic amino acid at this position in mouse C9 (Stanley & Herz, 1987). At position 395 a threonine rather than proline residue was found, con-

Table I: Nucleotide Sequence of Exon–Intron Junctions in the C9 Gene[a]

Exon	Nucleotide position	Exon intron boundaries
1	-81	.............ACCAGgtgagtca
2	82-187	tatcttagTTATG.............AAATGgtaagtgt
3	188-332	ttgtgcagTTTCG.............TACAGgtaatctt
4	333-479	gattgcagGCAGA.............CTATGgtgtgtat
5	480-627	tttctcagGGATC.............AAAGGgtctctgg
6	628-874	........CGAGA.............AGAAG.........
7	875-1115	caacacagGAAAA.............GAAAGgtatgcct
8	1116-1244	acaggcagGTGTT.............AGCTGgtgagtg
9	1245-1420	atatctagTAAAC.............AAAAAgtgagaac
10	1421-1650	........CTGTC.............GAAGG........
11	1651-	aaagagagATTGC.............

[a] Exon sequence is shown in upper case; intron sequence is shown in lower case. The exon–intron boundaries for exons 6 and 10 were not determined. Exon 1 starts in the 5′ untranslated region, and exon 11 ends in the 3′ noncoding region of the cDNA. Nucleotide numbers are for the corrected C9 sequence with an additional valine residue at position 293.

firming asparagine-393 (or 394 including the additional valine) as a site for attachment of N-linked oligosaccharides (Stanley et al., 1985). In each case the sequence of two human cDNA clones (unpublished data) and the genomic C9 sequence agree, suggesting that the discrepancies are a cloning artifact in the original cDNA clones.

Correlation between Exon–Intron Boundaries and Putative Protein Domains. At the amino terminus of human C9 are two short sequence motifs, each containing six cysteine residues, which are homologous to repeating sequence motifs in other proteins. Residues 18–77 correspond to a sequence found three times on each polypeptide chain of thrombospondin and twice in the complement components C7, C8α, and C8β. Residues 78–116 are highly homologous with a region repeated seven times in the LDL receptor which constitutes the apolipoprotein E/B binding site (the "class A" cysteine-rich seuence motif; Stanley et al., 1986). In neither case does the exon structure correlate with the boundaries of the sequence motif. The thrombospondin-like sequence is encoded by parts of exons 2 and 3 while the class A cysteine-rich motif is encoded by parts of exons 3 and 4 (Figure 2).

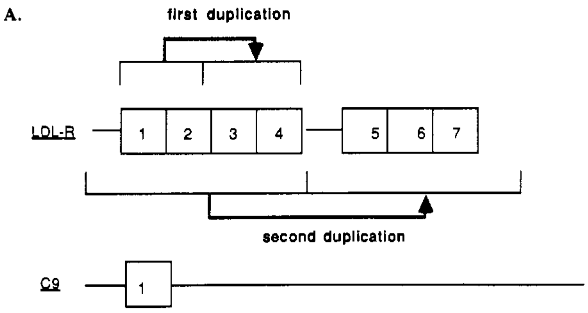
Proteolytic cleavage studies have revealed two hypersensitive regions in the molecule (Biesecker et al., 1982). One, near the center of the molecule, which is cleaved by α-thrombin and chymotrypsin is present in a region of low sequence conservation which might function as a loose connecting piece in the extension of C9 during membrane insertion (Stanley & Herz, 1987). A second site 146 amino acid residues from the carboxy terminus is preferentially cleaved by trypsin. After more extensive treatment with trypsin a second cleavage occurs close to the α-thrombin site to generate 18- and 20-kdalton fragments representing the central and carboxy-terminal portions of the molecule. It is likely that the central 18-kdalton fragment is identical with that remaining in membranes after extensive trypsin digestion of C9 inserted into biological membranes (Hammer et al., 1977), since almost all the polypeptide chain labeled by a membrane-restricted photoactivatable probe (Schäfer et al., 1987) is present in this region (Figure 1). Although these two regions of the molecule appear to be relatively stable to further proteolytic digestion, suggesting a compact globular structure, they are each encoded by three exons (exons 6–8 and 9–11 in Figure 1).

```
........tttagaccataccctttttttggcattagaaattctaatcattccaactgata
ttgtgcagTTTCGTTCAAGAAGCATTGAGGTCTTTGGACAATTTAATGGGAAAAGATGCAC
        heArgSerArgSerIleGluValPheGlyGluPheAspGlyLysArgCysThr
CGACGCTGTGGGAGACAGACGACAGTGTGTGCCCACAGAGCCCTGTGAGGATGCTGAGGAT
AspAlaValGlyAspArgArgGluCysValProThrGluProCysGluAspAspGluAspA
GACTGCGGAAATGACTTTCAATGCAGTACAGgtaatctttgtgcttgaattgctcattgtg
spCysGlyAspAspPheGluCysSerThrGl
gctttctgtgtgccttcctgaatgaaaagaaaaaaaaaagcatatgcttctggacacttgg
atttccaagcgtgaaaggccccaaaagaggccatctatatattctgtacaaagaactcaaa
tgtgtttttctgatacctcacctccagggttaaaatgcatttgataacttcagaaagaaag
agaaaatttgacattagatacattgagtctctcctgattttgattgcGCAGATGCATAAAG
                                                 yArgCysIleLysM
ATGCGACTTCGGTGTAATGGTGACAATGACTGCGGAGACTTTTCAGATGAGGATGATTGTG
etArgLeuArgCysAspGlyAspAspAspCysGlyAspPheSerAspGluAspAspCysGl
AAAGTGAGCCCCGTCCCCCCTGCAGAGACAGAGTGGTAGAAGAGTCTGAGCTGGCACGAAC
uSerGluProArgProProCysArgAspArgValValGluGluSerGluLeuAlaArgThr
AGCAGGCTATGGtgtgtattttacttgtactttttcagatgaaaatgagtgaaaatgatct
AlaGlyTyrGlu
ccatctctctcattgtgatagactggtagaagcatttgtgcgagggacataggtgata...
```

FIGURE 2: Nucleotide sequence of the region of the C9 gene encoding the class A cysteine-rich sequence motif (underlined). The derived amino acid sequence is shown below the nucleotide sequence of exons 3 and 4.

Correlation of Surface Features and Intron–Exon Boundaries. Surface features of C9 that have been mapped to sequences on the polypeptide chain include two sites of N-linked oligosaccharide attachment (Stanley et al., 1985), three sites of proteolytic enzyme cleavage (Stanley et al., 1985), and seven antibody epitopes (Stanley & Herz, 1987). In addition, turns in the predicted secondary structure would be expected to lie on the surface of the protein in most cases.

Of the 10 intron–exon boundaries 7 fall within predicted turn or coil regions in the predicted secondary structure of human C9, and 2 others are present at the end of predicted elements of secondary structure. Some splice junctions appear to be especially prominent surface features, like the junction between exons 8 and 9, which falls between residues 393 and 394. This is adjacent to the trypsin cleavage site before residue 392 and an N-linked oligosaccharide at residue 394 (Figure 1). The junction of exons 1 and 2 is also presumably an exposed feature since this is coincident with the polypeptide loop forming one of the epitopes of C9 antiserum (Stanley & Herz, 1987). The boundary of exons 3 and 4 is of particular interest since this occurs within the class A cysteine-rich se-



B.

	1	2	3	4	5	6	7	C9
1	x	.43	.68	.46	.53	.50	.52	.58
2		x	.43	.66	.34	.46	.31	.44
3			x	.51	.52	.42	.52	.47
4				x	.42	.52	.37	.46
5					x	.39	.33	.53
6						x	.47	.43
7							x	.45
C9								x

FIGURE 3: Correlation between the predicted structure of class A repeats. (A) Schematic representation of the duplication proposed to have occurred in the LDL receptor (see text). (B) Correlation coefficients relating pairs of class A cysteine-rich sequence motifs were calculated by averaging coefficients obtained from six physical parameters. Correlation coefficients between repeats related on the basis of the above model are underlined. Numbers 1–7 refer to the seven class A repeats of the LDL receptor numbered from the amino terminus.

quence motif. The interrupted amino acid in human C9 (residue 89) lies in precisely the same position as an insertion of two amino acids in one of the LDL receptor class A repeats (Yamamoto et al., 1984). This is therefore most likely a surface feature within the cysteine-rich domain.

Relationship between C9 and LDL Receptor Class A Sequence Motifs. In Figure 3 each class A sequence has been compared on the basis of physical parameters involved in protein folding (Argos, 1987). The correlation coefficients relating the sequences in Figure 3B are consistent with a pattern of successive duplications in the LDL receptor in which units 1 and 2 duplicate to form units 3 and 4 and then units 1–4 duplicate to generate units 5–8 (Figure 3A). Since the eighth unit is absent in the LDL receptor cDNA sequence, it must have been either deleted or lost by an alternative splicing event. On this scheme the first LDL receptor sequence is projected as the original sequence from which the others were derived, although a similarity cannot be discerned between units 1 and 2. This might reflect the conservation of a ligand binding site between dissimilar cysteine-rich domains as is found for the lectin binding site in wheat germ agglutinin (Wright, 1980). The class A sequence motif of C9 is most like the first repeat of the LDL receptor. A similar conclusion is suggested by the finding that both the C9 class A motif and the first motif of the LDL receptor are capable of binding calcium ions (Thielens et al., 1988; van Driel et al., 1987).

Discussion

There is good evidence in a number of proteins for structural or functional units being encoded by discrete exons with splice junctions coincident with the boundaries of the protein motif. It has been suggested that this structure could facilitate the rapid evolution of new protein designs by recombination between exons or their shuffling between genes (Gilbert, 1985, 1978; Blake, 1978, 1983). It is less clear how this gene structure arose, whether exons evolved as gene segments encoding motifs of a protein or whether exons evolved as random open-reading frames which were later trimmed to facilitate exon shuffling. A second, and independent, hypothesis is that splice junctions are located at the surface of a protein, where variations caused by recombination between exons is less likely to affect protein structure (Craik et al., 1982).

We have tested these two hypotheses in the case of human complement component C9, for which a lot of data concerning domain structure and surface topology are known. In general a good correlation between exon splice sites and surface features was observed, but none of the putative domains of the protein aligned with single exons. This was particularly surprising because C9 is one of a group of recently described mosaic proteins which contain cysteine-rich domains on discrete exons. While exons encoding cysteine-rich domains have been found fused together in both the LDL receptor class A motifs (Südhof et al., 1985a) and in class B (epidermal growth factor like) repeats in uromodulin (Pennica et al., 1987), there are no examples to our knowledge of introns interrupting these cysteine-rich motifs or of additional flanking sequences present in the same exon. Since the class A cysteine-rich sequence motif of C9 bears a high degree of homology to those of the LDL receptor, it was predicted that it would be present on a single exon since it was assumed that the two proteins inherited the common sequence by exon shuffling (Südhof et al., 1985b). The finding that the class A cysteine-rich sequence motif was in fact encoded by a part of two exons, with additional flanking sequences present in the exons and an interrupting intron sequence, suggests a more complex relationship between the two genes. While it is of course impossible to verify the evolutionary origin of the genes, it is clear that the class A sequence motif of the LDL receptor could not have been directly transferred from an acestral C9 gene by exon shuffling. Either the homologous amino acid segments in C9 and the LDL receptor are an example of convergent evolution, or substantial exon rearrangements must have occurred during their divergence from a common ancestral gene. In the latter case comparison of the two genes is instructive since they might indicate stages in the development of an exon-encoded structural domain of a protein.

At least two steps are required to convert the C9 exon pattern into that found in the LDL receptor, or vice versa. If the LDL receptor is considered as the older gene, in accord with the concept that the original exons evolved as discrete structural domains, then the C9 exon–intron arrangement could only have occurred by intron insertion in the center of the class A sequence and loss of introns to generate the flanking sequences in exons 3 and 4. Alternatively, if the C9 gene is regarded as being closer to the ancestral class A gene, then the rearrangement of introns involves an intron loss in the center of the sequence and a relocation of the flanking introns to the boundary of the protein domain (Figure 4). Intron insertion, which has to be postulated if the LDL receptor is the older gene, has been shown to occur only in a few cases (Rogers et al., 1985). Furthermore, analysis of the sequence (Figure 3) and Ca^{2+} binding properties of the class A repeats in C9 and the LDL receptor supports the view that the sequence in C9 is most like the first repeat in the LDL receptor, which in turn is likely to be the ancestor of the other repeats. Thus the C9 gene is likely to be at least as old as the LDL receptor gene.

In the C9 gene the arrangement of exons with respect to putative protein domains supports a random mechanism for the evolution of exons as open-reading frame segments of

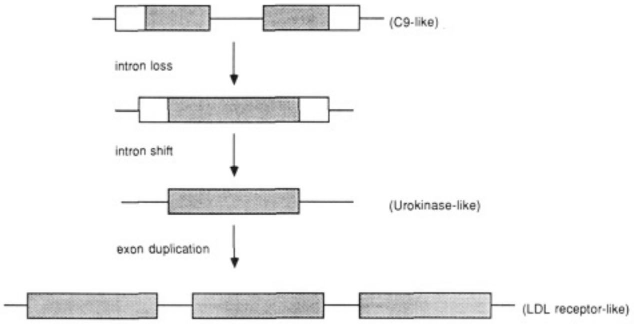


FIGURE 4: Schematic representation of intron rearrangement in genes coding for the cysteine-rich domains. Boxes represent exon sequences; lines represent intron sequences. Cysteine-rich regions are shaded.

primordial RNA (Darnell & Doolittle, 1986), rather than the creation of exons as discrete units coding for the protein structural motifs. Assuming that the class A motif of the LDL receptor is indeed derived from that of C9, then it is likely that the exon rearrangement occurred before or during duplication since otherwise multiple changes would be necessary. One possibility is that the changes were linked to the mechanism of duplication of the exon which would be necessary in order to form proteins like the LDL receptor with multiple copies of the motif.

The steps of a possible intron rearrangement are shown schematically in Figure 4. Between the exon arrangement of genes like C9 and those like the LDL receptor is one in which the cysteine-rich region is encoded by a single exon, but in which duplication has not yet occurred. A possible example of this is found in the urokinase gene which has a single class B repeat which is precisely bounded by an exon (Riccio et al., 1985).

In summary, while exon splice sites correlate well with surface features of C9, they do not correlate with the boundaries of the protein domains. In the case of the class A cysteine-rich domain, substantial intron rearrangements would be required in order to generate the unit for duplication and exon shuffling which has occurred in the LDL receptor gene. The gene organization of exon units encoding cysteine-rich domains in C9, urokinase, and the LDL receptor may therefore represent different stages in the evolution of an exon-encoded protein motif in which the exons were fused and trimmed until a unit suitable for shuffling was obtained.

Acknowledgments

We are grateful to J. Dickson for technical assistance and to Drs. A. M. Frischauf, J. Collins, and H. Haymerle for kindly providing the human genomic libraries.

Registry No. C9, 80295-59-6.

References

Argos, P. (1987) *J. Mol. Biol. 193*, 385–396.

Biesecker, G., Gerard, C., & Hugli, T. E. (1982) *J. Biol. Chem. 257*, 2584–2590.

Blake, C. C. F. (1978) *Nature (London) 273*, 267.

Blake, C. C. F. (1983) *Nature (London) 306*, 535–537.

Church, G. M., & Gilbert, W. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 1991–1995.

Craik, C. S., Sprang, S., Fletterick, R., & Rutter, W. J. (1982) *Nature (London) 299*, 180–182.

Darnell, J. E., & Doolittle, W. F. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 1271–1275.

DiScipio, R. G., Gehring, M. R., Podack, E. R., Kan, C. C., Hugli, T. E., & Fey, G. H. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 7298–7302.

DiScipio, R. G., Chakravarti, D. N., Müller-Eberhard, H. J., & Fey, G. H. (1988) *J. Biol. Chem. 263*, 549–560.

Doolittle, W. F. (1985) *Trends Biochem. Sci. (Pers. Ed.) 10*, 233–237.

Drenth, J., Low, B. W., Richardson, J. S., & Wright, C. (1980) *J. Biol. Chem. 255*, 2652–2655.

Feinberg, A. P., & Vogelstein, B. (1983) *Anal. Biochem. 132*, 6–13.

Frischauf, A. M., Lehrach, H., Poustka, A., & Murray, N. (1983) *J. Mol. Biol. 170*, 827–842.

Gilbert, W. (1978) *Nature (London) 271*, 501.

Gilbert, W. (1985) *Science (Washington, D.C.) 228*, 823–824.

Hammer, C. H., Shin, M. L., Abramovitz, A. S., & Mayer, M. M. (1977) *J. Immunol. 119*, 1–8.

Hohn, B., & Collins, J. (1980) *Gene 11*, 291–298.

Lawler, J., & Hynes, R. O. (1986) *J. Cell. Biol. 103*, 1635–1648.

Maniatis, T., Hardison, R. C., Lacy, E., Lauer, J., O'Connell, C., Quon, D., Sim, G. K., & Estratiadis, A. (1978) *Cell (Cambridge, Mass.) 15*, 687–701.

Maniatis, T., Fritsch, E. F., & Sambrook, J. (1982) *Molecular Cloning Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

Mount, S. M. (1982) *Nucleic Acids Res. 10*, 459–472.

Norrander, J., Kempe, T., & Messing, J. (1983) *Gene 26*, 101–106.

Pennica, D., Kohr, W. J., Kuang, W.-J., Glaister, D., Aggarwal, B. B., Chen, E. Y., & Goeddel, D. V. (1987) *Science (Washington, D.C.) 236*, 83–88.

Podack, E. R., & Tschopp, J. (1982a) *J. Biol. Chem. 257*, 15204–15212.

Podack, E. R., & Tschopp, J. (1982b) *Proc. Natl. Acad. Sci. U.S.A. 79*, 574–578.

Poustka, A., Rackwitz, H. R., Frischauf, A. M., Hohn, B., & Lehrach, H. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 4129–4133.

Rackwitz, H. R., Zehetner, G., Frischauf, A. M., & Lehrach, H. (1984) *Gene 30*, 195–200.

Riccio, A., Grimaldi, G., Verde, P., Sebastio, P., Boast, S., & Blasi, F. (1985) *Nucleic Acids Res. 13*, 2759–2771.

Rigby, P. W. J., Dieckmann, M., Rhodes, C., & Berg, P. (1977) *J. Mol. Biol. 113*, 237–251.

Rogers, J. (1985) *Nature (London) 315*, 458–459.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 5463–5467.

Sanger, F., Coulson, A. R., Barrell, B. G., Smith, A. J. M., & Roe, B. A. (1980) *J. Mol. Biol. 143*, 161–178.

Schäfer, S., Amiguet, P., & Tschopp, J. (1987) *Complement 4*, 220.

Southern, E. M. (1975) *J. Mol. Biol. 98*, 503–517.

Stanley, K. K., Page, M., Campbell, A. K., & Luzio, J. P. (1986) *Mol. Immunol. 23*, 451–458.

Stanley, K. K., & Herz, J. (1987) *EMBO J. 6*, 1951–1957.

Stanley, K. K., Kocher, H. P., Luzio, J. P., Jackson, P., & Tschopp, J. (1985) *EMBO J. 4*, 375–382.

Südhof, T. C., Goldstein, J. L., Brown, M. S., & Russell, D. W. (1985a) *Science (Washington, D.C.) 228*, 815–822.

Südhof, T. C., Russell, D. W., Goldstein, J. L., & Brown, M. S. (1985b) *Science (Washington, D.C.) 228*, 893–895.

Thielens, N. M., Lohner, K., & Esser, A. F. (1988) *J. Biol. Chem. 263*, 6665-6670.

van Driel, I. R., Goldstein, J. L., Südhof, T. C., & Brown, M. S. (1987) *J. Biol. Chem. 262*, 17443-17449.

Wright, C. S. (1980) *J. Mol. Biol. 141*, 267-291.

Wyman, A. R., Wertman, K. F., Barker, D., Helms, C., & Petri, W. H. (1986) *Gene 49*, 263-271.

Yamamoto, T., Davis, C. G., Brown, M. S., Schneider, W. J., Casey, M. L., Goldstein, J. L., & Russell, D. W. (1984) *Cell (Cambridge, Mass.) 39*, 27-38.

Human H1 Histone Gene Promoter CCAAT Box Binding Protein HiNF-B Is a Mosaic Factor[†]

Andre J. van Wijnen, Robert F. Massung, Janet L. Stein,* and Gary S. Stein*

Department of Cell Biology, University of Massachusetts Medical School, Worcester, Massachusetts 01655

Received February 23, 1988; Revised Manuscript Received April 13, 1988

ABSTRACT: Vertebrate histone gene promoters in many cases contain an upstream element, 5′dCCAAT, that has been implicated in modulating the efficiency of transcription of a broad spectrum of genes. We have previously isolated a nuclear factor (HiNF-B) that binds specifically to the CCAAT element of a cell cycle regulated human H1 histone gene. This factor shows similarities with other CCAAT box binding proteins in that it recognizes the same sequence but shows a distinct chromatographic behavior. In the present study, we have employed the gel retardation assay to demonstrate that HiNF-B is a cell cycle independent DNA binding protein that is conserved in both human and mouse cells. Using a series of reconstitution experiments with partially purified HiNF-B fractions, we show that this factor requires association of at least two components for site-specific binding. The composite structure of HiNF-B suggests that binding of at least some CCAAT elements in vertebrates may require cooperative interaction of CCAAT box binding proteins with other factors.

Sequence-specific DNA binding proteins play key roles in the transcriptional regulation of eukaryotic gene expression (Dynan & Tjian, 1985). Vertebrate promoter DNA sequences acting as binding sites for such proteins include the TATA element (Parker & Topol, 1984), the CCAAT element (Graves et al., 1986), the GGGCGG element (Dynan & Tjian, 1983), and the ATTTGCAT element (Singh et al., 1986). The CCAAT element can be found in a variety of vertebrate promoters and has been implicated in the regulation of a number of genes, such as the Herpes Simplex Virus thymidine kinase gene (Jones et al., 1985) and mouse α-globin (Mellon et al., 1981) and β-globin genes (Dierks et al., 1983). This transcriptional element has been shown to act as a recognition site for CCAAT box binding proteins (McKnight & Tjian, 1986; Cohen et al., 1986; Myers et al., 1986; van Wijnen et al., 1988). In addition, the CCAAT element has been implicated in the initiation of adenovirus DNA replication (Jones et al., 1987).

The eukaryote-specific H1 and core (H2A, H2B, H3, and H4) histone genes encode a set of proteins that are essential for maintaining the integrity of the eukaryotic genome and are encoded in multiple clusters in the vertebrate genome. Coordinate regulation of this multigene family occurs in tight conjunction with DNA replication, and control is exerted at both transcriptional and posttranscriptional levels [reviewed in Stein et al. (1984) and Schumperli (1986)]. Vertebrate histone genes are constitutively transcribed at a basal rate in actively dividing cells, and at the onset of S phase in the cell division cycle, a 3-5-fold transient enhancement of the transcription rate occurs (Plumb et al., 1983; Heintz et al., 1983; Sittman et al., 1983; Artishevsky et al., 1987; Baumbach et al., 1987).

The transcriptional regulation of human H1 and core histone genes is mediated by a battery of cis-acting elements (Sierra et al., 1983; Kroeger et al., 1987; Helms et al., 1987; Sive et al., 1986; Dailey et al., 1986), some of which are histone specific and others that are shared by a broad spectrum of human promoters. Several of these histone promoter elements coincide with DNA/protein interaction sites in vitro (van Wijnen et al., 1987, 1988; Dailey et al., 1986; Sive & Roeder, 1986) and with regions of chromatin hypersensitive to DNaseI, MNase, and S1 nuclease (Pauli et al., 1988; Moreno et al., 1986; Chrysogelos et al., 1985). Moreover, our laboratory has established in vivo sites of DNA/protein interaction in the proximal promoter region of two cell cycle regulated core (H3 and H4) histone genes. Both the H4 and H3 histone genes have two major regions of in vivo DNA/protein interactions (site I and site II) that overlap in vitro protein binding sites (Pauli et al., 1987; unpublished data). Sequences encompassing the entire H4 promoter site II, which spans both a (distal) histone-specific element and a (proximal) TATA box, are required and sufficient for accurate H4 histone mRNA transcription initiation in vivo (Kroeger et al., 1987). Although various transcriptional elements and associated nuclear factors have been identified, the role of these in the coordinate increase in histone mRNA synthesis in early S phase is not clear.

The proximal promoter region of a cell cycle regulated human H1 histone gene contains two target sites for the binding of nuclear factors in vitro (van Wijnen et al., 1988). A distal domain is bound by HiNF-A, a factor that binds to at least two other (H3 and H4) histone gene promoters (van Wijnen et al., 1987). A proximal domain (analogous in position to the H4 promoter site II) functions as a binding site for HiNF-B, a novel CCAAT box binding protein. In this work, we present evidence that both factors have a cell cycle

[†] This work was supported by grants from the National Institutes of Health (GM 32010 and GM 32381), the National Science Foundation (DCB 8896116), and the March of Dimes Birth Defects Foundation (1-813).

0006-2960/88/0427-6534$01.50/0 © 1988 American Chemical Society