Journal of Cellular Biochemistry 33:185–190 (1987)
Proteases in Biological Control and Biotechnology 169–174

Structure and Evolution of the Human Genes Encoding Protein C and Coagulation Factor IX

George L. Long

Division of Molecular and Cell Biology, Lilly Research Laboratories, Indianapolis, Indiana 46285

Human protein C is a vitamin K-dependent plasma protein that serves as a feedback down-regulator of the coagulation cascade by specifically degrading the protein cofactors VIIIa and Va. The protein C precursor consists of the following domains: leader peptide, "gla" region, two epidermal growth factor segments, and the activation peptide/serine protease. Comparison of amino acid sequences reveals that protein C and factor IX are homologous. A comparison of the genes for protein C and factor IX shows that all seven of the introns within the protein coding regions are in identical positions and correspond to protein structure-function domain boundries. However, the base compositions of the two genes (coding and noncoding regions) are remarkably different: ~60% guanine + cytosine (G + C) for protein C versus ~40% G + C for factor IX. One possible explanation for this phenomenon is that the factor IX gene (located on the X chromosome) has undergone extensive deoxycytosine methylation and subsequent spontaneous deamination mutagenesis, resulting in a net C to thymine (and G to adenine) transition. This would suggest that the protein C gene may represent a more primitive form of the gene duplication precursor.

Key words: **protein C, factor IX, coagulation, methylation, methyl cytosine, intron, serine protease, evolution, mutation, gene structure**

Protein C is a blood plasma protein involved in the regulation of hemostastis [1]. At the site of clot formation, circulating inactive protein C zymogen precursor is converted to an active serine protease by limited thrombin cleavage. Activated protein C specifically degrades two nonenzymatic protein cofactors in the coagulation cascade, Va and VIIIa, thereby serving as an on-demand, feedback down-regulator of both the intrinsic and extrinsic pathways of coagulation. Human protein C has been purified and partially characterized [2].

George L. Long's present address is Department of Biochemistry, School of Medicine, University of Vermont, Burlington, VT 05405.

Received May 16, 1986; accepted September 15, 1986.

Recently, the entire amino acid sequence of human protein C precursor has been reported [3], based upon the cDNA sequence. The precursor schematically consists of five distinct domains (Fig. 1): leader peptide (amino acids −42 to −1), a γ-carboxyglutamate (Gla) segment (aa 1–45), two epidermal growth factor domains (aa46–91, 92–137), and a classical trypsin-like serine protease region (aa158–419).

Human protein C precursor undergoes extensive post-translational modification including vitamin K-dependent carboxylation of nine glutamyl residues, hydroxylation of an aspartyl residue, glycosylation (23% weight), disulfide bond formation, and limited proteolysis. Circulating inactive protein C consists predominantly of a "light" (Mr = 21,000) and a "heavy" chain (Mr = 41,000) joined by disulfide bridging. The circulating two-chain molecule is generated by cleavage of a linking Lys-Arg dipeptide from the internal portion of the precursor single-chain protein [3].

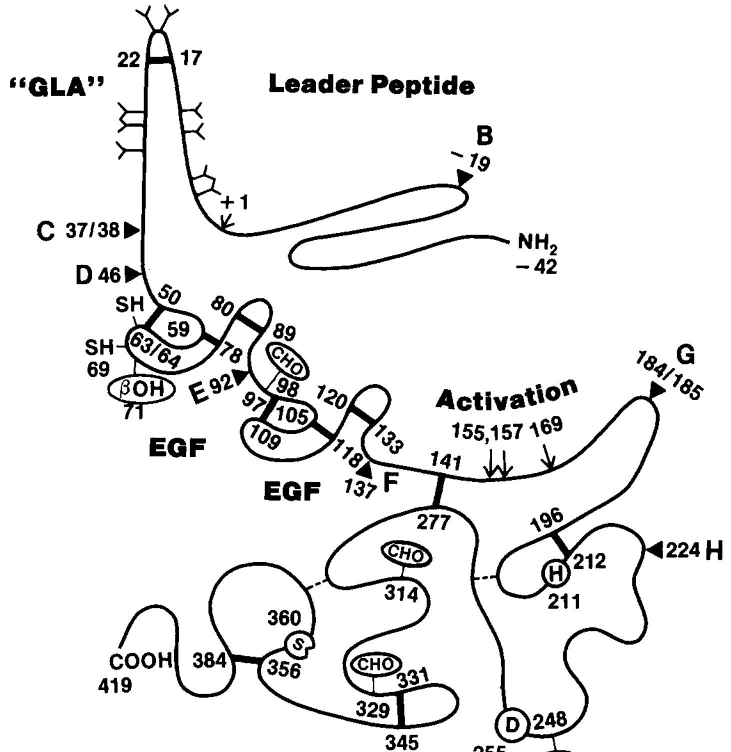

Fig. 1. Schematic representation of protein C structure. Single-chain precursor is represented by the thin curving line. Numbers refer to animo acid positions in the precursor protein. See the text and reference [3] for a description of the noted protein domains. Proposed positions of the disulfide bridges (thick bars), β-hydroxyaspartate (β-OH), γ-carboxyglutamate (Y), and carbohydrate attachment (CHO) in the mature protein are also shown, and are based upon homology with related proteins as discussed in ref. [3]. The catalytic triad (Ser, His, Asp) of the serine protease domain are represented by the circled letters S, H, and D, respectively. Known proteolytic cleavage sites resulting in two-chain, activated protein C are shown with arrows. The corresponding positions of introns B-H (see Tables I and II) in the gene are portrayed by closed triangles (▲) and are described in detail in ref [7].

COMPARISON OF PROTEIN C AND FACTOR IX

Protein C shares sequence homology with other vitamin K-dependent coagulation factors, including factors VII, IX, X, and prothrombin [4,5]. For example, human protein C has sequence homology throughout (34% identical amino acids overall) when compared with human coagulation factor IX [3], and consequently they share the same protein domains. Because of the structural closeness of the two proteins, a comparison of the genes for human protein C [6,7] and factor IX [8,9] was performed. Both genes contain seven intervening segments (introns) in the protein-coding portions of the mRNA precursors; and *all* seven of the introns occur at identical positions in the two genes [7]. Figure 1 shows the position of the introns relative to the protein sequence. The position of the introns correlates well with the protein's proposed structural domain boundaries. These results are consistent with the hypothesis put forth by Gilbert several years ago [10], that introns may bound coding segments for distinct protein structural domains and may participate in the genetic reshuffling of common domains among many different proteins.

In contrast to their positional identity, the introns for protein C and factor IX show no similarity in size (Table I). Furthermore, the base composition and sequence of the two genes are dissimilar. Figure 2 is a comparison of base composition (guanine + cytosine) for the seven positionally common introns. Surprisingly, all of the protein C introns are richer in G and C bases than are the factor IX counterparts. A comparison of 5′ with 3′ ends of individual introns (data not presented) shows no consistent difference in G + C content and cannot account for the results presented in Figure 2.

Also shown in Figure 2 is a comparison of the protein-coding portions of the two genes. The G + C content of the coding segments of the protein C gene is 21% greater than that for factor IX. Table II compares the coding regions for the two genes in terms of codon usage. The codon usage patterns for the two genes are very different. Table II shows a general cytosine to thymine (T) and guanine to adenine (A) transition when comparing protein C codons to those of factor IX. A distinct difference in codon usage is seen even for amino acids predominantly found in highly homologous regions of the proteins. For example, glutamic acid and cysteine are major components of the "Gla" and epidermal growth factor domains, which possess 59% and 36% amino acid sequence identity, respectively [3]. Table III further

TABLE I. Size of Introns for Human Protein C and Factor IX

Intron[a]	Protein C	Factor IX
B	1.25 kb	5.76 kb
C	1.55	0.19
D	0.09	4.06
E	0.10	7.52
F	2.84	2.57
G	1.15	10.06
H	1.12	0.67

[a] Data taken from references [7,8]. Intron names are assigned with B representing the 5′ most extreme intron within the coding portion of the gene, C representing the next intron downstream toward the 3′ end, etc. The gene for protein C also contains an intron (A) in the 5′ noncoding region that is ~ 1.44 kb in size [7].

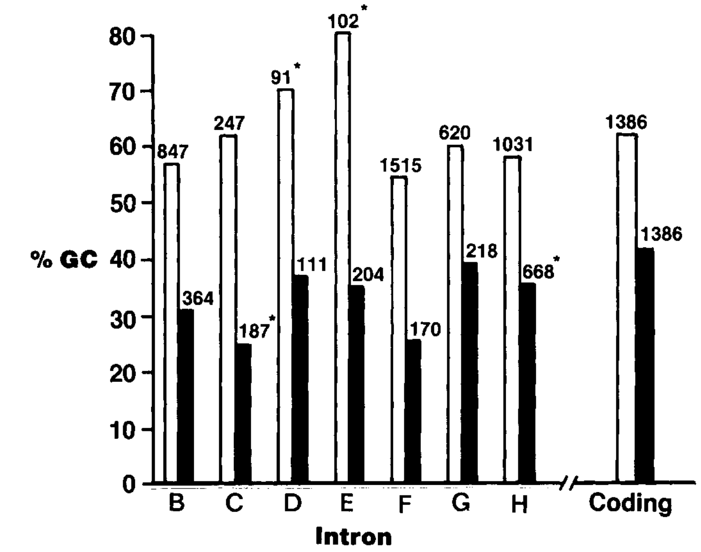

Fig. 2. Comparison of the guanine + cytosine base content for the human protein C and factor IX genes. The percent G + C for bounding regions of introns within the coding portions of human protein C (open bar) and factor IX (solid bar) genes are shown. Also shown is the G + C content for the two coding sequences. Numbers above each bar show the number of nucleotides included. Entire intron compositions are noted by asterisks (*). Intron names are as described in Table I. Compostional data are taken from references [7, 8].

TABLE II. Human Codon Usage (462 Codons) For Protein C and Factor IX[a]

	PC	IX		PC	IX		PC	IX		PC	IX
TTT:Phe	0	12	TCT:Ser	1	6	TAT:Tyr	1	11	TGT:Cys	7	19
TTC:Phe	15	9	TCC:Ser	7	5	TAC:Tyr	7	5	TGC:Cys	17	5
TTA:Leu	0	6	TCA:Ser	2	6	TAA:Stop	0	1	TGA:Stop	0	0
TTG:Leu	3	3	TCG:Ser	2	0	TAG:Stop	1	0	TGG:Trp	15	7
CTT:Leu	5	9	CCT:Pro	4	4	CAT:His	2	6	CGT:Arg	3	1
CTC:Leu	16	5	CCC:Pro	11	3	CAC:His	16	4	CGC:Arg	10	1
CTA:Leu	1	2	CCA:Pro	1	8	CAA:Gln	3	7	CGA:Arg	2	6
CTG:Leu	24	3	CCG:Pro	4	0	CAG:Gin	12	7	CGG:Arg	8	3
ATT:Ile	4	17	ACT:Thr	0	12	AAT:Asn	5	15	AGT:Ser	2	7
ATC:Ile	12	7	ACC:Thr	10	7	AAC:Asn	8	17	AGC:Ser	22	3
ATA:Ile	2	1	ACA:Thr	5	10	AAA:Lys	3	12	AGA:Arg	2	8
ATG:Met	8	6	ACG:Thr	3	1	AAG:Lys	21	16	AGG:Arg	3	1
GTT:Val	1	22	GCT:Ala	1	10	GAT:Asp	5	12	GGT:Gly	2	8
GTC:Val	11	3	GCC:Ala	15	5	GAC:Asp	25	7	GGC:Gly	20	8
GTA:Val	1	5	GCA:Ala	5	8	GAA:Glu	4	33	GGA:Gly	3	15
GTG:Val	16	7	GCG:Ala	3	1	GAG:Glu	30	10	GGG:Gly	10	4

[a]Data taken from references [7,8] for human protein C (PC) and factor IX (IX), respectively.

TABLE III. Comparison of Base Composition Versus Codon Position for Human Protein C and Factor IX

	1st Position			2nd Position			3rd Position		
Base	PC	IX	Δ[a]	PC	IX	Δ[a]	PC	IX	Δ[a]
T	78	95	17	119	117	2	43	171	128
C	122	69	53	74	86	12	222	94	128
A	110	140	30	143	163	20	34	128	94
G	152	158	6	126	96	30	161	69	92

[a]Difference in the No. of the particular base between human protein C (PC) and factor IX (IX). First position: $\Sigma\Delta = 106$; second position: $\Sigma\Delta = 64$; third position: $\Sigma\Delta = 442$.

compares base composition with codon usage in regard to position in the codons. As expected from a consideration of the genetic code redundancy, the greatest difference in base frequency between the two genes is in the third position, and the smallest difference is in the second position [11]. The largest change is a C ↔ T transition in the third position. Possibly related to this is the fact that the most abundant base for both genes in the first position is guanine.

DISCUSSION

A comparison of the intron positions for human protein C and factor IX genes shows that they are identical, which suggests that the two genes have a close evolutionary relationship. However, a comparison of intron size and gene base composition shows that the genes are significantly divergent. Protein C gene contains ~20% more G + C bases than does factor IX. This disparity exists also in the coding regions of the two genes.

A recently popular and successful technique for detecting desired cDNA clones involves the use of a unique oligonucleotide "guessmer" probe [12,13]. The probe sequence is based upon a partial amino acid sequence and assumed most-preferred codons for the protein in question. The probe is sufficiently long enough (40–60 nucleotides) to compensate for any infrequent mismatches. The results reported here should serve as a point of caution for individuals contemplating the use of the "guessmer" technique. Protein C and factor IX are two highly homologous proteins, both synthesized in the liver at approximately equal levels. However, the base composition (Fig. 2) and codon usage (Table II) are significantly different.

No clear explanation for the compositional divergence of these two evolutionary closely related genes exists. One possibility is based upon observed high mutation rates in bacteriophage [14], bacteria [15,16], *Neurospora* [17], and eukaryotes [18] of cytosine to thymine (and indirectly G → A on the opposite strand) via methylation and subsequent deamination. In animals the predominant (90–95%) site of methylation is at the CpG dinucleotide [19]. The report of Selker and Stevens [17] on the mutation of *Neurospora* 5S RNA genes is particularly significant. Two of the genes, ζ (zeta) and η (eta), lie close together and are apparently the result of a tandem duplication. Within the duplicated 794 nucleotide regions there are 113 base differences, all of which are C-T or G-A differences. Comparison with other *Neurospora* 5S RNA genes also shows that the mutations were polarized C → T, G → A, and the ζ and η regions are heavily methylated. The fact that within the repeated regions there are no transversion mutations (G-C, G-T, A-C, or A-T) suggests to the authors that the duplication was a recent evolutionary event, followed by the accumulation of transition mutations via methylation and mutation. In the case of the protein C and factor IX genes, methylation of cytosine residues followed by mutations to thymine could result in the compositional differences reported here. Consistent with this hypothesis are the observations that the most common transition (C-T) occurs in the third codon position, which is the most permissive position (ie, no amino acid change), and is most frequently followed by a guanine base (first position). This situation reflects the predominant site of eukaryotic DNA methylation: CpG (TpG resulting in the mutation product) [19]. One consequence of this hypothesis is that the gene containing the greater amount of C + G bases (protein C) can be considered as more closely representing the gene duplication precursor.

One additional observation that may be related to the above hypothesis is the chromosomal location of the genes for protein C and factor IX. The human protein C gene is located on chromosome 2 (unpublished results, Susan S. Naylor and G.L. Long), whereas factor IX resides on the X chromosome [20]. DNA methylation appears to play an important role in X-chromosome inactivation [21].

Although the above hypothesis may help to explain the process by which the two genes have diverged, the question of why the proposed precursor form of the genes and the present form of the protein C gene are so G + C rich (total mammalian DNA is ~60% A + T) is unanswered.

ACKNOWLEDGMENTS

The author wishes to thank R.H. Baltz for bringing reference [11] to the author's attention and Ms. Sherry Pike for excellent secretarial assistance.

REFERENCES

1. Esmon CT, editor: "Seminars in Thrombosis and Hemostasis: Protein C," Vol 10. pp. 109–172, New York: Thieme-Stratton, Inc., 1984.
2. Kisiel W: J Clin Invest 64:761, 1979.
3. Beckman RJ, Schmidt RJ, Santerre RF, Plutzky G, Crabtree GR, Long GL: Nucleic Acid Res 13:5233, 1985.
4. Fernlund P, Stenflo J: J Biol Chem 257:12170, 1982.
5. Stenflo J, Fernlund P: J Biol Chem 257:12180, 1982.
6. Foster DC, Yoshitake S, Davie EW: Proc Natl Acad Sci USA 82:4673, 1985.
7. Plutzky J, Hoskins J, Long GL, Crabtree GR: Proc Natl Acad Sci USA 83:546, 1986.
8. Anson DS, Choo KH, Rees DJG, Giannelli F, Gould K, Huddleston JA, Brownlee GG: EMBO Journal 3:1053, 1984.
9. Yoshitake S, Schach BG, Foster DC, Davie EW, Kurachi K: Biochemistry 24:3736, 1985.
10. Gilbert W: Nature 271:501, 1978.
11. Woese CR: Naturwissenschaften 60:447, 1973.
12. Anderson A, Kingston IB: Proc Natl Acad Sci USA 80:6838, 1983.
13. Pennica D, Nedwin GE, Hayflick JS, Seeburg PH, Deynick R, Palladine MA, Kohr WJ, Aggarwal BB, Goeddel DV: Nature 312:724, 1984.
14. Baltz RH, Bingham PM, Drake JW: Proc Natl Acad Sci USA 73:1269, 1976.
15. Coulondre C, Miller JH, Farabaugh PJ, Gilbert W: Nature 274:755, 1978.
16. Duncan BK, Weiss B: J Bacteriol 151:750, 1982.
17. Selker EU, Stevens JN: Proc Natl Acad Sci USA 82:8114, 1985.
18. Taylor JH: "DNA Methylation and Cellular Differentiation." New York: Springer-Verlag, 1984.
19. Doscocil J, Sorm F: Biochim Biophys Acta 55:953, 1962.
20. Camerino J, Grzeschik KH, Jaye M, De La Salle H, Tolstoshow: Proc Natl Acad Sci USA 81:498, 1984.
21. Gartler SM, Riggs AD: Annu Rev Genet 17:155, 1983.