J Mol Evol (2008) 66:185–196
DOI 10.1007/s00239-008-9074-8

Genomic Evidence for a Simpler Clotting Scheme in Jawless Vertebrates

Russell F. Doolittle · Yong Jiang · Justin Nand

Received: 23 October 2007 / Accepted: 25 January 2008 / Published online: 19 February 2008

Abstract Mammalian blood clotting involves numerous components, most of which are the result of gene duplications that occurred early in vertebrate evolution and after the divergence of protochordates. As such, the genomes of the jawless fish (hagfish and lamprey) offer the best possibility for finding systems that might have a reduced set of the many clotting factors observed in higher vertebrates. The most straightforward way of inventorying these factors may be through whole genome sequencing. In this regard, the NCBI Trace database (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi) for the lamprey (*Petromyzon marinus*) contains more than 18 million raw DNA sequences determined by whole-genome shotgun methodology. The data are estimated to be about sixfold redundant, indicating that coverage is sufficiently complete to permit judgments about the presence or absence of particular genes. A search for 20 proteins whose sequences were determined prior to the trace database study found all 20. A subsequent search for specified coagulation factors revealed a lamprey system with a smaller number of components than is found in other vertebrates in that factors V and VIII seem to be represented by a single gene, and factor IX, which is ordinarily a cofactor of factor VIII, is not present. Fortuitously, after the completion of the survey of the Trace database, a draft assembly based on the same database was posted. The draft assembly allowed many of the identified Trace fragments to be linked into longer sequences that fully support the conclusion that lampreys have a simpler clotting scheme compared with other vertebrates. The data are also consistent with the hypothesis that a whole-genome duplication or other large scale block duplication occurred after the divergence of jawless fish from other vertebrates and allowed the simultaneous appearance of a second set of two functionally paired proteins in the vertebrate clotting scheme.

Keywords Blood clotting · Lamprey genome · Gene duplications

Introduction

Blood coagulation is known to follow a similar scheme in all vertebrates, the culminating event being the thrombin-catalyzed conversion of fibrinogen into fibrin. Interest in how the process evolved to yield the complex system that occurs in mammals has been longstanding, not only because every component seems essential, but also because some of the factors—like factors IX and X—depend on others—like factors VIII and V—for their activity. It has long been appreciated that a series of different gene duplications gave rise to many of the factors, and it was hoped that studies on early diverging vertebrates, especially jawless fish, might reveal a simpler process as it existed in earlier times.

Electronic supplementary material The online version of this article (doi:10.1007/s00239-008-9074-8) contains supplementary material, which is available to authorized users.

R. F. Doolittle (✉) · Y. Jiang · J. Nand
Department of Chemistry & Biochemistry, University of California, San Diego, La Jolla, CA 92093-0314, USA
e-mail: rdoolittle@ucsd.edu

R. F. Doolittle · Y. Jiang · J. Nand
Department of Molecular Biology, University of California, San Diego, La Jolla, CA 92093-0314, USA

Present Address:
Y. Jiang
Moores Cancer Center, University of California, San Diego, La Jolla, CA 92093, USA

However, it has proved difficult to assess whether the entire constellation of factors leading to thrombin generation in mammals is present in these early-diverging vertebrates. For one thing, demonstrating the presence or absence of a particular factor by biochemical assay in such creatures is handicapped by "species specificity," which confounds classical measurements that depend on the use of standardized mammalian protein reagents or genetically defective plasmas. Although prothrombin and fibrinogen have long ago been purified and the presence of tissue factor in lampreys demonstrated biochemically (Doolittle et al. 1962; Doolittle and Surgenor 1962), it has not been realistic to attempt the purification of other coagulation factors, several of which are present in only minute amounts in mammalian plasma. Some clotting factors have been cloned from lampreys and hagfish (Strong et al. 1985; Bohonus et al. 1986; Wang et al. 1989; Banfield and MacGillivray 1992; Pan and Doolittle 1992) and several more from later-diverging teleosts like zebrafish (Jagadeeswaran et al. 2000; Sheehan et al. 2001; Hanumanthaiah et al. 2002) and puffer fish (Davidson et al. 2003a). Now, with the advent of whole-genome sequencing (WGS), it has become feasible to reconstruct the ensemble of clotting proteins that occurs in early diverging vertebrates.

In this regard, recent studies based on the complete genome sequence of the puffer fish, *Fugu rubripes*, revealed that genes for most known clotting proteins—including all of the vitamin K-dependent factors and the critical cofactor proteins, factors V and VIII—are present (Davidson et al. 2003a, b; Jiang and Doolittle 2003). For the most part, only genes for a few of the more peripheral coagulation factors, like factors XI and XII, were not found (Jiang and Doolittle 2003). The puffer fish is a teleost, however, and it would be of much greater interest if such a study could be conducted on one of the jawless fish—lamprey or hagfish—which diverged 50 million to 100 million years earlier (Carroll 1988) and would be more likely to have a simpler clotting system. None of the principal clotting factors is found in the genome of the protochordate *Ciona intestinales* (Jiang and Doolittle 2003).

At present, a complete genome sequence is not yet available for either lamprey or hagfish. However, a "trace database" maintained by the NCBI and EBI includes the lamprey among its numerous holdings. Trace databases are uncurated collections of DNA sequences, mostly determined by random shotgun methods at major sequencing centers around the world. In the case of the lamprey (*Petromyzon marinus*), the November 2006 collection contained 18,787,613 machine-generated "reads," or "traces," most between 300 and 1000 "letters," amounting to 14,640,144,063 nucleotides (nt). The lamprey genome is estimated to contain between 1.6 billion and 2.2 billion nt (Gregory 2005), suggesting that the average redundancy in the database is about six- or sevenfold. The current study began with a computer search of the Trace database in an effort to identify various coagulation factors.

The degree of coverage notwithstanding, determining whether a gene is present or not in a Trace database is much more challenging than is the case when a complete and assembled genome is at hand. In the case of the puffer fish, for example, data were available in the form of 12,381 scaffolds that had been assembled from the original DNA fragments. The average scaffold size was such that in most cases all the exons of a given gene were present on a single assembly. In the lamprey Trace database, exons or parts of them were scattered over the approximately 18 million entries, minimizing chances of verification by exons from the same gene being linked. However, these collections do contain a small fraction of EST sequences (determined from cDNA), and these are usually longer and often provide overlaps. Additionally, many of the Trace sequences are available as "mate-pairs," sequences determined from the two different "ends" of inserts in vectors, occasionally permitting estimates about positional neighborliness in the genome.

Recently, after the completion of the original set of searches, a partial assembly of the same lamprey data became available. Obviously, assembly data greatly facilitate the reconstruction of gene sequences, and it was a straightforward matter to match up the initial findings with the partially assembled data. The average size of the "contigs" (or supercontigs) that link together fragments from the original database is about 7000, which is an order of magnitude greater resolution than was the case for the original "traces." As a result, in several cases entire genes are found on single entities.

The most serious complication for the present study stems from most of the gene products of interest having resulted from gene duplications that took place about the same time as the appearance of vertebrates, aggravating the problem of distinguishing orthologues from highly similar paralogues. In particular, in a previous study (Jiang and Doolittle 2003) we had found that factors V and VIII in the puffer fish are only 42% identical to their human counterparts, and in either species the two factors themselves are 38% identical, indicating that the gene duplication giving rise to these two factors occurred not very long before the divergence of bony fish and the line leading to mammals and, perhaps, after the divergence of the jawless fish.

The cases of factors V and VIII are even more problematic because of their being members of the ferroxidase family of proteins, which includes ceruloplasmin and hephaestin, themselves composed of three major domains that are the result of (tandem) duplications (Hellman and Gitlin 2002). In other vertebrates, factors V and VIII are distinguishable from hephaestin and ceruloplasmin in that they have "discoidin" domains at their carboxyl termini (sometimes called "fac5–8 C" domains).

Methods

The publicly accessible lamprey Trace Archive data were downloaded from the NCBI web site (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi), after which they were reformatted into a form compatible with previously described software (Doolittle 1987). Each DNA entry was translated into amino acid sequences corresponding to all six frames, which were then treated separately. In addition, BLAST software (Altschul et al. 1997) was downloaded from the NCBI web site; tblastn was used extensively, with the translated sequences being searched against the raw DNA data.

Searches for exact matches at least 30 codons in length were conducted on 20 proteins whose sequences had been determined before the inception of genome projects. These searches were conducted both with in-house software and with tblastn. The same dual strategy was employed in the search for coagulation factors. Verification of matches was obtained by using BLAST to search the NCBI nonredundant protein database, both before and after concatenating various translated traces to form longer sequences. Validation of a proper assignment was achieved, or not, by the nature of the top hits in this reverse search. Phylogenetic reconstructions were made by a distance-matrix method (Feng and Doolittle 1996), a parsimony procedure (Doolittle and Feng 1990), and a neighbor-joining method (Saitou and Nei 1987); the software for the latter was downloaded from PHYLIP (Felsenstein 1989). In the case of the distance-based trees, the BLOSUM62 substitution matrix was used (Henikoff and Henikoff 1996). Trees were drawn on the PHYLODENDRON web site, http://www.iubio.bio.indiana.edu/treeapp/treeprint-form.html.

Mate-pairs A part of the Trace archiving strategy involves determining the sequences of both "ends" of every DNA fragment, and in several cases reciprocal matches were made that allowed neighboring exons to be linked. The Trace ID numbers of all mate-pairs were compared with those of all candidates in a particular venue (i.e., factor V–VIII study, for one, and separately for the vitamin K-dependent protease study). Additionally, a perl script was written that allowed translated sequences of identified mate-pairs to be used in conjunction with blastp to learn what kind of domains they might contain. The Trace ID numbers of all fragments matching the various targets are provided in tabular form in the Supplementary Material.

The draft assembly data were produced by the Genome Sequencing Center at Washington University School of Medicine in St. Louis and were obtained from ftp://genome.wustl.edu/pub/petromyzon_marinus. The web site notes that the original WGS data were obtained from a specimen provided by M. Bronner-Fraser, which was sequenced to a total 5.9× whole-genome coverage. Because the draft assembly is based on the same Trace collection used in our study, it was a relatively straightforward matter to screen the posted list of "reads.placed" to find the contigs and supercontigs on which our original "hit-groups" are situated. In many cases it was possible to link hit-groups into consecutive strings, and in several instances (especially in the cases of the vitamin K-dependent proteases) entire genes were found to be encompassed on single supercontigs. In some cases, however, the hits are on small supercontigs or were terminal segments in the "wrong" direction with regard to expected neighbors and remain unlinked "orphans."

Results

Completeness of the Lamprey Trace Database

The extent of coverage of the lamprey Trace database was estimated by searching for exact matches for 20 lamprey genes (or proteins) whose sequences had been determined before the WGS lamprey project was undertaken. These included a wide assortment of proteins not involved in blood clotting, as well as data for fibrinogen chains and a fragment of prothrombin. All 20 were positively identified by multiple "hits" (Table 1), demonstrating that the lamprey Trace database was sufficiently redundant to make judgments about the presence or absence of particular genes.

Factors V and VIII

In other vertebrates, factors V and VIII are composed of three A domains (~340 amino acids each) and two carboxyl-terminal discoidin domains (Fig. 1). There is also a long, extended region between the second and the third A domains, denoted the B domain, which varies greatly between factor V and factor VIII, as well as between species. In humans, the B segment of factor VIII extends for 1035 residues, and the corresponding region in factor V amounts to 940 residues. With the exception of a short segment at the carboxyl-terminal end, there is no detectable sequence similarity between the human versions of factors V and VIII in these regions; searches of these two B

Table 1 Proteins whose sequences were known in advance of the genome project and that were used to test for completeness of coverage

Query protein	Length	No. hits[a]	Trace ID[b]
Apolipoprotein 1	105	8	745174549
Apolipoprotein 2	191	21	1200323037
Cytochrome c	104	72	1188802510
Globin V	149	98	813242897
Insulin	57	11	1199723668
Enolase (frag)	97	35	1470242083
Lactate dehydrogenase	334	92	745175264
Plasma albumin	551 (120)[c]	146	1222023964
Plasmin (frags)	327	11	813242595
Rhodopsin	353	42	1171186792
Vitellogenin	1823 (150)[c]	8	194145636
Superoxide dismutase	144	49	1210986068
Fibrinogen			
γ chain	408	46	813242447
β chain	479	92	1161196149
α chain	961	16	1159459473
α2 chain	641	28	1171833083
Prothrombin (frag)	82	29	1198677997
Unspecified serpin	448 (150)[c]	15	1161339639
CD45 phosphatase	1000 (100)[c]	10	1207673272
Netrin	199	74	1169718195

[a] Perfect matches for 30-residue segments

[b] The Trace ID given is for only one representative “read”

[c] Only a part of sequence used for searching

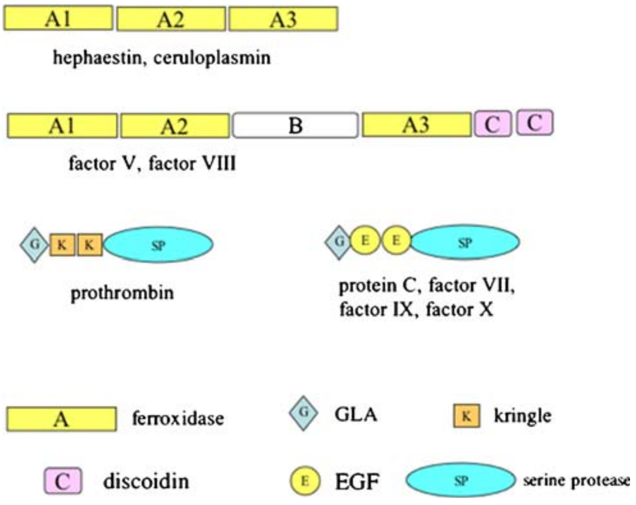

Fig. 1 Cartoon showing domain structures of the principal proteins dealt with in this study

domains against the lamprey Trace database did not yield any significant hits.[1]

When sequences of the three A domains of human factors V and VIII were used as queries in a search of the lamprey Trace database, a large number of “hits” was obtained. Unexpectedly, when subjected to “back-searching” (reciprocal matching) against NCBI data protein databases, the majority of these were more similar to hephaestin and ceruloplasmin than to factor V or VIII. Nonetheless, a long list of candidate “traces” was compiled, and those portions showing any detectable similarity to human factors V and VIII were retrieved.

All told, searches of the six A domains (three each from human factors V and VIII) identified approximately 257 “hits” in the lamprey trace database. When these were compared with each other, many were found to contain the same (or virtually the same) inserts, and the number was subsequently reduced to 50 “hit-groups,” consistent with an approximate fivefold redundancy. There was a wide range, however, the number of identical inserts in the various groups ranged from 0 to 29.

Of the 50 unique hits, fewer than a dozen resembled factors V or VIII more than or as much as they did hephaestin or ceruloplasmin.[2] Concerned that the initial search might have been compromised by the long query sequences that are inconsistent with exon-length targets, we redid the entire exercise on an exon-by-exon basis, using human exon sequences as queries.

Exon-intron distribution of hephaestin-family genes In humans, the hephaestin and ceruloplasmin genes each have 18 introns that demarcate 19 exons (Syed et al. 2002; Daimon et al. 1995). Factors V and VIII have 18 and 19 introns in the corresponding regions, respectively (Cripe et al. 1992). The exon sizes for the four human proteins range from 30 to 85 codons in length (Fig. 2). The fact that the majority of pairs of exons in human hephaestin and ceruloplasmin genes have slightly different lengths than occur in factors V and VIII served as a useful, if tentative, property for assigning exons from the lamprey Trace database to one set or the other. It was also helpful that the exon boundaries for the three homologous A domains differ in all four of these proteins.

[1] The search of the factor V B domain actually detected a large number of almost-perfect tandem 27-nt repeats in the lamprey Trace database. Although both human and mouse factor V have 30 imperfect copies of these tandem nine-amino acid repeats, none occurs in the factor V sequences of chicken or puffer fish; this coincidental but perhaps chance similarity does not bear directly on the problem at hand.

[2] One of the Trace sequences (G52 in Fig. 2; Trace ID 1446326143) exhibited a remarkable 77% identity to a 34-residue segment of human factor VIII (26 identities among the 34 residues) and may be a contaminant. The corresponding region from puffer fish is only 44% identical to the human segment. The same region from chicken is coincidentally 77% identical to the human sequence, but the eight differences are not the same as observed in the lamprey sequence.

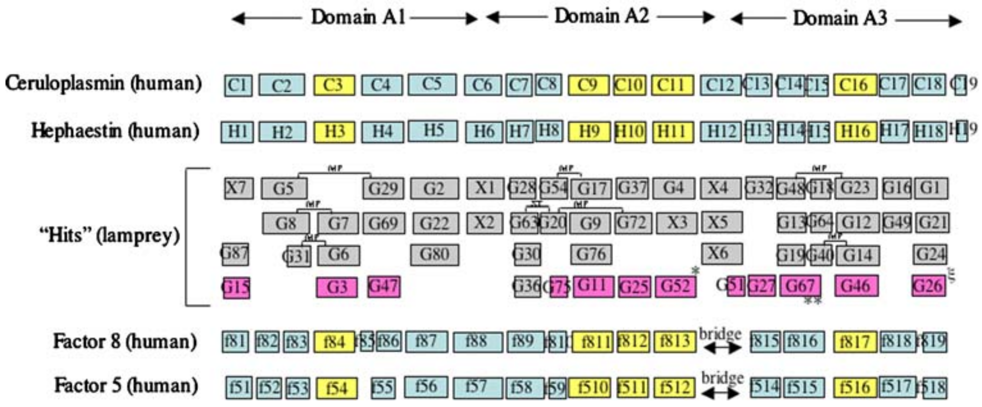
Domain A1
Domain A2
Domain A3
Ceruloplasmin (human)
C1 C2 C3 C4 C5 C6 C7 C8 C9 C10 C11 C12 C13 C14 C15 C16 C17 C18 C19
Hephaestin (human)
H1 H2 H3 H4 H5 H6 H7 H8 H9 H10 H11 H12 H13 H14 H15 H16 H17 H18 H19
"Hits" (lamprey)
MP
ST
X7 G5 G29 G2 X1 G28 G54 G17 G37 G4 X4 G32 G48 G18 G23 G16 G1
G8 G7 G69 G22 X2 G63 G20 G9 G72 X3 X5 G13 G64 G12 G49 G21
G87 G31 G6 G80 G30 G76 X6 G19 G40 G14 G24
G15 G3 G47 G36 G75 G11 G25 G52 G51 G27 G67 G46 G26
*
**
ξ
Factor 8 (human)
f81 f82 f83 f84 f85 f86 f87 f88 f89 f810 f811 f812 f813 bridge f815 f816 f817 f818 f819
Factor 5 (human)
f51 f52 f53 f54 f55 f56 f57 f58 f59 f510 f511 f512 bridge f514 f515 f516 f517 f518

Fig. 2 Schematic alignment of exons from human genes for hephaestin, ceruloplasmin, and factors V and VIII (light blue) compared with corresponding "hits" identified in the lamprey Trace database (gray, more similar to hephaestin or ceruloplasmin; pink, more similar to factor V or VIII; orange, borderline between two groups). Thirteen introns occur at equivalent locations in all four genes. Four others are offset by only 12–14 codons. A single intron is unique to factor VIII; another occurs in factors V and VIII but not in hephaestin and ceruloplasmin. The five sets of exons colored yellow have coincident boundaries in all four human genes. The membership (Trace ID numbers) of all "hit-groups" is provided as Supplementary Material. ST: exons from same trace. MP: mate-pairs. G67 (marked **) has intron boundaries like factors V/VIII but has a sequence more similar to ceruloplasmin. G26 (marked ξ) has intron boundaries like hephaestin/ceruloplasmin but a sequence more similar to factor V. G52 (marked *) has an anomalously high resemblance to human factor VIII.[2]

Although no additional hits were found (i.e., beyond the original 257) when exons from factors V and VIII were searched individually, several more hits were added when the searches were conducted with the individual exons of human hephaestin and ceruloplasmin. Hits found only with hephaestin or ceruloplasmin query sequences have been designated differently (X1–7 in Fig. 2) in order to distinguish them from those found with factors V–VIII. The various "hit groups" were assigned to either the hephaestin-ceruloplasmin family or the factor V–VIII family on the basis of sequence similarity and exon sizes. In seven cases it was possible to link segments using "mate-pairs"(MP in Fig. 2). In one instance, two exons were found in a single trace, out of frame and separated by a small intron.

An analysis of matching segments make it clear that, at a minimum, there are four genes in lamprey that belong to the hephaestin-ceruloplasmin-factor V/VIII family, only one of which seems closer to the FV/VIII side of the family (Fig. 2). The question to be decided was, Are genes for *both* factors V and VIII present, or is there only a single progenitor (preduplication) gene? If there is only one gene of the factor V–VIII type, then there would reasonably have to be an "extra"(homologue) gene for either hephaestin or ceruloplasmin. It is known that, at the very least, lamprey blood plasma contains a blue protein with all the properties of ceruloplasmin (J. Gitlin, unpublished data).

The posting of the draft assembly of the lamprey genome made it possible to screen the list of "traces" provided and to identify the supercontigs on which the various hit-groups occur. In this regard, the same factor V–VIII hit-groups shown in Fig. 2 are presented in Fig. 3, except in their relative positions on 16 supercontigs. All told, 37 of the 57 hit-groups initially used to reconstruct these entities were found in the draft assembly. Three others could be linked to particular supercontigs by the use of mate-pairs observed in the Trace database. Hit-groups not yet found in the assembly are listed in their original positions across the bottom of Fig. 3 (except G51 was shifted to a better matching position). Three other hepaestin-related exons were found in the same set of supercontigs (labeled H10, H17 and discoidin in Fig. 3). All the assignments are consistent with there being a maximum of four genes belonging to the hephaestin-ceruloplasmin-factor V–VIII family, only one of which is closer to factors V or VIII.

To reinforce the point, a series of linked segments from the A3 domains of all four putative proteins was compared with the corresponding regions of hephaestin, ceruloplasmin, and factors V and VIII from humans (Fig. 4A). Most significantly, the intron positions in three of the lamprey genes are the same as occur in the hephaestin-ceruloplasmin branch; in only one of the lamprey sequences do the introns coincide with those in factors V–VIII (Fig. 4A). Phylogenetic trees based on these sequences were entirely corroborative (Figs. 4B and C).

Vitamin K-Dependent Proteases

With regard to the vitamin K-dependent proteases, the main challenge was distinguishing them from the profusion of other paralogous serine proteases and assigning them to their proper orthologue category. As in the case of factors V and VIII, searches were conducted across a range of

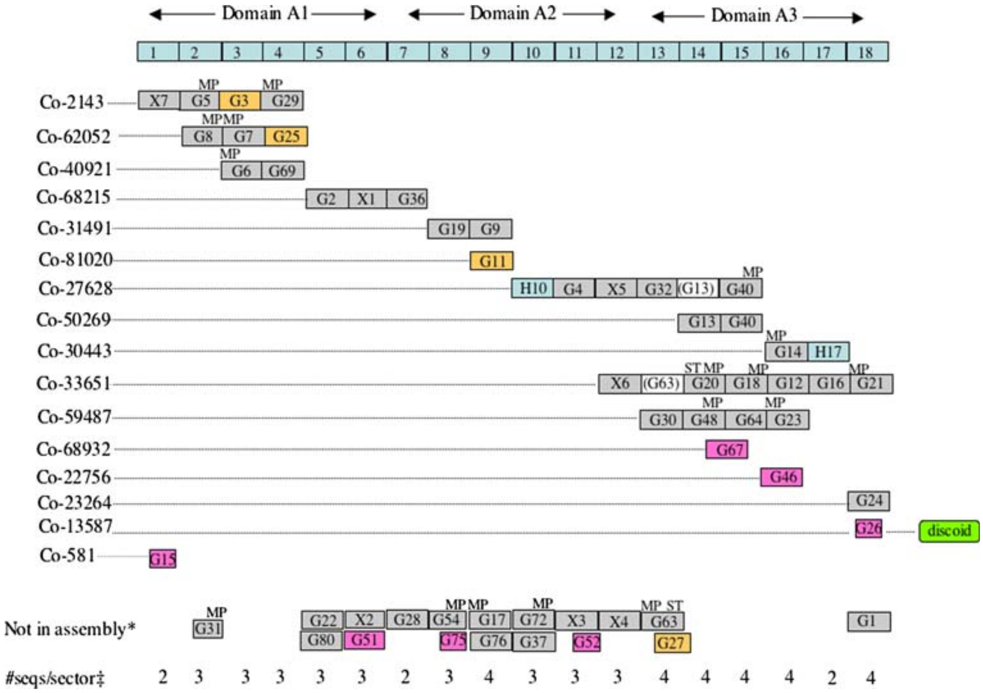

Fig. 3 Sixteen supercontigs that include 36 "hit-groups" for blood clotting factors V and VIII and the related proteins hephaestin and ceruloplasmin (to be compared with Fig. 2, text). Eighteen of the 57 hit-groups not found in the draft assembly are shown across the bottom of the figure. The 18 sectors (top, blue boxes) correspond to exons found in human hephaestin. The number of sequences found for each sector is shown at the bottom (seq/sect). The colors correspond to those employed in Fig. 2. Open boxes represent regions where sequences are missing in middle regions of contigs (NNNN regions) but where sequences are available elsewhere, either in other contigs or in mate-pairs in the Trace database

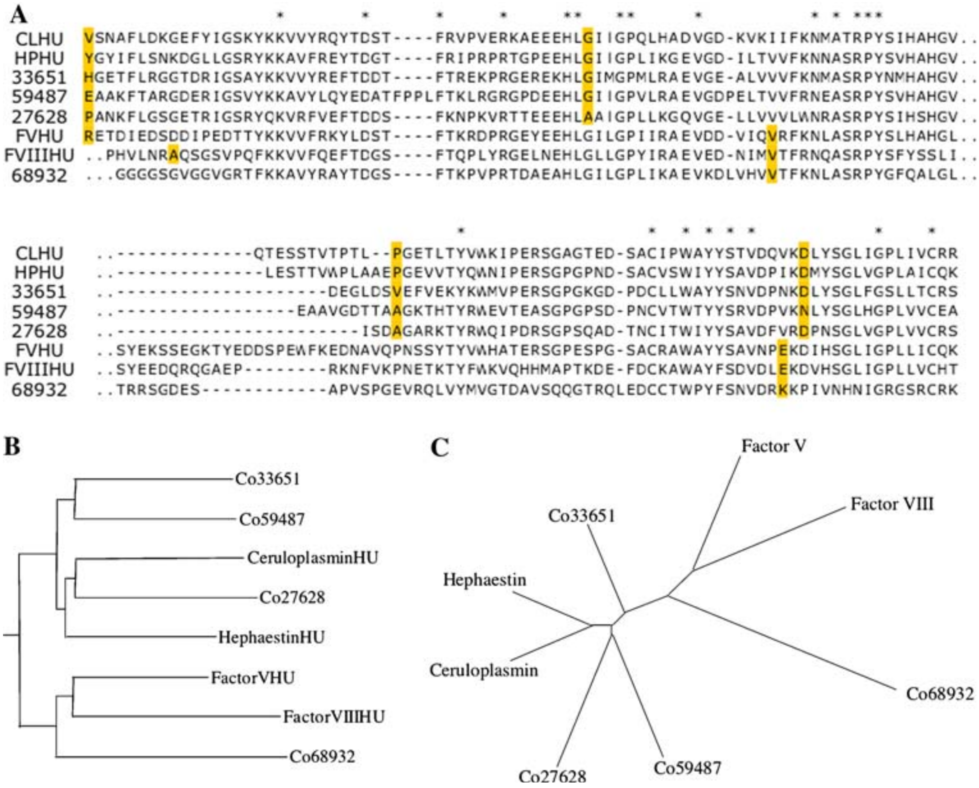
A
CLHU VSNAFLDKGEFYIGSKYKKVVYRQYTDST----FRVPVERKAEEEHLGIIGPQLHADVGD-KVKIIFKNMATRPYSIHAHGV..
HPHU YGYIFLSNKDGLLGSRYKKAVFREYTDGT----FRIPRPRTGPEEHLGIIGPLIKGEVGD-ILTVVFKNNASRPYSVHAHGV..
33651 HGETFLRGGTDRIGSAYKKVVYREFTDDT----FTREKPRGEREKHLGIMGPMLRAEVGE-ALVVVFKNMASRPYNMHAHGV..
59487 EAAKFTARGDERIGSVYKKAVYLQYEDATFPPLFTKLRGRGPDEEHLGIIGPVLRAEVGDPELTVVFRNEASRPYSVHAHGV..
27628 PANKFLGSGETRIGSRYQKVRFVEFTDDS----FKNPKVRTTEEEHLAAIGPLLKGQVGE-LLVVVLWNRASRPYSIHSHGV..
FVHU RETDIEDSDDIPEDTTYKKVVFRKYLDST----FTKRDPRGEYEEHLGILGPIIRAEVDD-VIQVRFKNLASRPYSLHAHGL..
FVIIIHU ..PHVLNRAQSGSVPQFKKVVFQEFTDGS----FTQPLYRGELNEHLGLLGPYIRAEVED-NIMVTFRNQASRPYSFYSSLI..
68932 ...GGGGSGVGGVGRTFKKAVYRAYTDGS----FTKPVPRTDAEAHLGILGPLIKAEVKDLVHVVTFKNLASRPYGFQALGL..
CLHU ..-------------QTESSTVTPTL--PGETLTYVWKIPERSGAGTED-SACIPWAYYSTVDQVKDLYSGLIGPLIVCRR
HPHU ..--------------LESTTVWPLAAEPGEVVTYQWNIPERSGPGPND-SACVSWIYYSAVDPIKDMYSGLVGPLAICQK
33651 ..-------------------DEGLDSVEFVEKYKWMVPERSGPGKGD-PDCLLWAYYSNVDPNKDLYSGLFGSLLTCRS
59487 ..-----------------EAAVGDTTAAGKTHTYRWEVTEASGPGPSD-PNCVTWTYYSRVDPVKNLYSGLHGPLVVCEA
27628 ..----------------------ISDAGARKTYRWQIPDRSGPSQAD-TNCITWIYYSAVDFVRDPNSGLVGPLVVCRS
FVHU ..SYEKSSEGKTYEDDSPEWFKEDNAVQPNSSYTYVWHATERSGPESPG-SACRAWAYYSAVNPEKDIHSGLIGPLLICQK
FVIIIHU ..SYEEDQRQGAEP--------RKNFVKPNETKTYFWKVQHHMAPTKDE-FDCKAWAYFSDVDLEKDVHSGLIGPLLVCHT
68932 ..TRRSGDES------------APVSPGEVRQLVYMVGTDAVSQQGTRQLEDCCTWPYFSNVDRKKPIVNHNIGRGSRCRK
B
Co33651
Co59487
CeruloplasminHU
Co27628
HephaestinHU
FactorVHU
FactorVIIIHU
Co68932
C
Factor V
Factor VIII
Co33651
Hephaestin
Ceruloplasmin
Co68932
Co27628
Co59487

Fig. 4 (**A**) Sequence alignment of sectors 13–15 and amino-terminal portions of sector 16 (Fig. 3), where parts of all four lamprey genes can be compared directly. CLHU, human ceruloplasmin; HPHU, human hephaestin; FVHU, human factor V; FVIIIHU, human factor VIII. The linked groups have been named for their left-most contig (Fig. 3). Asterisks denote completely conserved residues. Orange bars denote intron positions. Linked groups 27628, 33651, and 59487 have intron positions that are coincident with the hephaestin-ceruloplasmin type; linked group 68932 has its introns in exactly the same place as human factors V and VIII genes. (**B**) Phylogenetic tree (Doolittle and Feng 1990) calculated from the alignment shown in A. (**C**) An unrooted neighbor-joining tree based on the same alignment (Saitou and Nei 1987; Felsenstein 1989)

sequence lengths, from the exon level to full length. In this regard, the distribution of introns varies greatly among genes for known vitamin K-dependent proteins (Fig. 5). Also, distinctions made on the basis of accessory domains—in this case the GLA (vitamin K-dependent), kringle, and EGF domains—were not immediately helpful because of their being genetically separated by introns from the main body of sequence encoding the protease. In one case, however, a mate-pair linkage was established between a GLA domain and a kringle, and in a few other

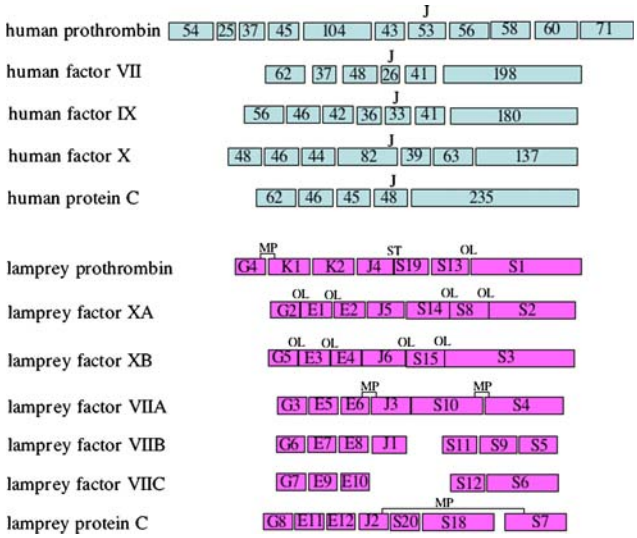

Fig. 5 Schematic alignment of exons from human genes for prothrombin, factors VII, IX, and X, and protein C (light blue; numbers denote size of exons in codons) compared with corresponding "hits" identified in the lamprey Trace database (letter-number combinations denote "hit-groups"; the complete list of ID numbers for all members of each hit group is provided as Supplementary Material). MP, mate-pairs; OL, overlapping fragments; ST, exons from same trace

cases EST sequences (from cDNA) in the database allowed domain connections to be made (Fig. 5).

GLA domain inventory An iterative search for sequences corresponding to GLA domains of human vitamin K-dependent proteases (prothrombin, factors VII, IX, and X, and protein C) and two nonproteases (proteins S and Z) resulted in 68 "hits," which, after analysis for redundancy, reduced to eight unique sequences (allowing for a few single-amino acid differences). This compares with 16 GLA domains[3] found in the (nonredundant) genome of the puffer fish (Jiang and Doolittle 2003) and 4 in the sea squirt (Jiang and Doolittle 2003; Kulman et al. 2006). If it is presumed that the lamprey genome coverage is sufficiently complete at this point, and that all related sequences have been identified, then the eight GLA sequences place an upper bound on the number of vitamin K-dependent proteins in the lamprey.

In mammals, GLA domains are also associated with proteins not involved in blood clotting, including bone proteins (Price et al. 1976), a cell growth potentiating factor (Manfioletti et al. 1993), and certain proline-rich proteins (Kulman et al. 1997). The GLA sequences involved with bone are quite dissimilar and may not have been picked up by our searches. In any event, the jawless fish do not have calcified bone and probably do not have the GLA-containing proteins associated with bone in other vertebrates. A search of the Trace database with mammalian Matrix GLA Protein and Bone GLA protein sequences (Laize et al. 2005) did not turn up any credible hits. In passing, it can be noted that in the puffer fish, only 10 of the 16 GLA domains are associated with the relevant proteases. Of the remaining six, one each occurs in proteins S and Z and the remaining four with proteins not involved in clotting.

Kringle and EGF domains Prothrombin is unique among the vitamin K-dependent proteases in that it has two kringle domains following the GLA domain; the other proteases in this group have two EGF domains at that location (Fig. 1). It was possible to identify the first prothrombin kringle by a mate-pair trace. A second was assigned initially on the basis of similarity to kringles in human and hagfish prothrombins. The overabundance of EGF domains in the Trace database precluded making unequivocal assignments merely on the basis of resemblance to human orthologues. However, EST sequences (from cDNA) for two related gene products resembling factor X included two EGF domains. It was also possible to link another EGF domain with a putative factor VII by a mate-pair (Fig. 5). The remaining EGF domains were assigned initially on the basis of scoring matches with human or puffer fish counterparts, whichever gave the higher score (Fig. 5).

Serine protease domains Several fragments for prothrombin, including a relatively long EST fragment, were immediately put in place by exact matching with a sequence previously determined with lamprey cDNA (Pan 1992). Some others were aided by resemblances to a published cDNA sequence for prothrombin from hagfish (Banfield and MacGillivray 1992). EST sequences were also helpful in linking peripheral domains to the main bodies of two putative factor X sequences. Coincidentally, these two entities had been cloned (cDNA) more than 10 years ago in the laboratory of S. Sommers (personal communication of unpublished material); the cDNA sequences are virtually identical to the sequences reconstructed from the Trace fragments. The strong resemblance between these two putative factor X sequences (>60% identical) implies a duplication that occurred well after the divergence of lampreys from other vertebrates. One of the gene products, denoted factor XB in Fig. 5, lacks a key residue in the activation sequence and ought not be active as a protease.

The bulk of the serine protease domain for a putative protein C was accounted for by four trace fragments, two of which were linked by mate-pairing (Fig. 5). The reconstructed sequence was 47% identical to the protease domain from human protein C. Similarly, the reconstructed

[3] Initially we reported that there were 19 GLA domains in the puffer fish (Jiang and Doolittle 2003), but some redundant sequences have apparently been removed from updated versions of that database, and we now count 16.

sequence for a factor VII was based on the similarity of various trace fragments to factors VII from human and fish. As it happens, there was evidence that more than one factor VII sequence is present in the lamprey, just as is the case in puffer fish (Davidson et al. 2003a, b; Jiang and Doolittle 2003). Three different trace sequences containing the cleavage activation point retrieved factor VII when blasted against the NCBI nonredundant database. In line with this finding, several other traces were found that best matched factor VII (Fig. 5). Because none of these could be linked to neighbors, phylogenetic trees were computed from an alignment of carboxy-terminal segments only (Figs. 6A and B). Notably, all three putative factor VII sequences, arbitrarily denoted A, B, and C, clustered together with human and puffer fish factor VII . One of the putative genes contains a nine-codon insertion in the active-site region and may be inactive.

When the draft assembly data became available, each of the 38 hit-groups used in making the initial reconstructions was looked for on the list of reads.placed; EGF domains aside, all but 3 were found in the assembly, including 7 of the 8 GLA-domains identified above. Five of these could be linked up with other hit-groups to complete full genes (Fig. 7). A sixth GLA was linked to an EGF domain, and the seventh was an “orphan”; the partially reconstructed factors VIIB and VIIC were limited to orphan supercontigs. Not unexpectedly, eight of the arbitrarily chosen EGF domains were replaced by other EGF domains in the assembly (Fig. 7).

Significantly, the full-length sequences of the protease portions of prothrombin, protein C, and factors VII, XA, and XB were exactly the same as the sequences reconstructed from Trace fragments. A phylogenetic tree calculated from an alignment of 20 full-length serine protease sequences, including lamprey prothrombin, protein C, two factors X, and factor VII, showed all the lamprey factors to be clustered with mammalian counterparts (Fig. 8). As in the case of the Trace database, no evidence was found in the draft assembly for a factor IX, despite exhaustive searching with factor IX sequences from a wide variety of species.

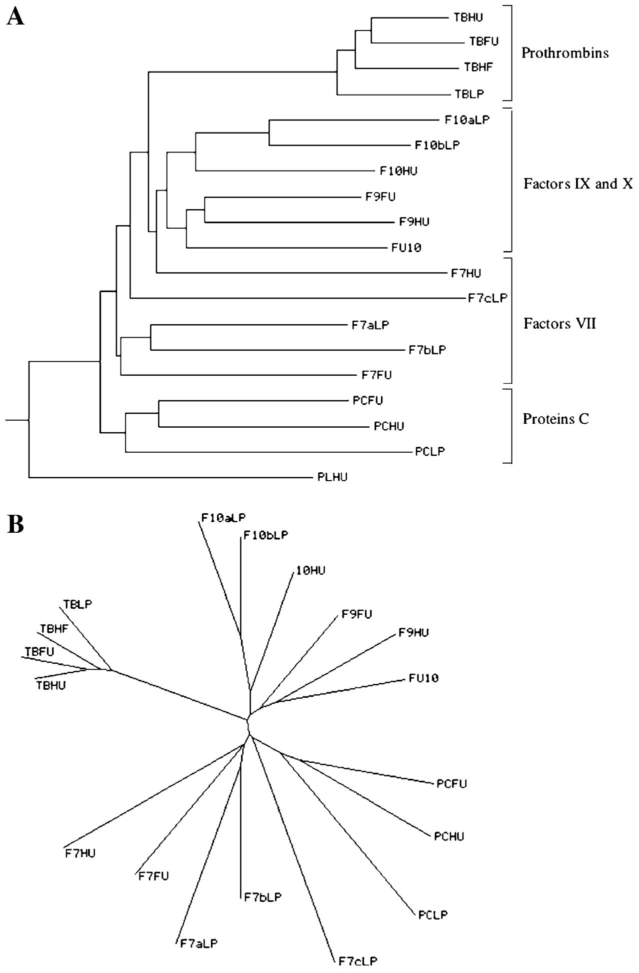

Fig. 6 Phylogenetic trees based on carboxy-terminal segments of seven putative vitamin K-dependent proteases found in the lamprey (LP) genome compared with counterparts from human (HU), puffer fish (FU), and hagfish (HF). TB, thrombin; PC, protein C; F7, factor VII; F9, factor IX; F10, factor X. (**A**) A tree generated by distance matrix method (Feng and Doolittle 1996). (**B**) An unrooted neighbor-joining tree based on the same alignment (Saitou and Nei 1987; Felsenstein 1989)

Other Coagulation Factors

Searches were conducted for two other vitamin K-dependent factors that are not proteases. A few marginal matches were found for portions of protein S, which in other vertebrates is composed of two laminin domains and four EGF domains besides its terminal GLA domain. The large number of EGF domain sequences in the trace database made positive identification difficult. As for the GLA domain, the sequence in human protein S is most similar to GLA hit-group 6, the only one of the eight GLA hit-groups not present in the assembly (Fig. 7). The case for protein S in the lamprey genome remains equivocal.

Nor was it possible to identify any reasonable candidates for protein Z, a fast-changing vitamin K-dependent factor that contains a nonfunctional protease portion (Ichinose et al. 1990). In contrast, strong matches for plasminogen (Trace ID 1184095239) and tissue plasminogen activator (Trace ID 1483701246) were inadvertently uncovered during the searches for vitamin K-dependent proteases.

We were unable to demonstrate the presence of a tissue factor (TF) sequence—another fast-changing entity—in either the trace database or the assembly, even though tissue factor has been demonstrated biochemically in lamprey tissues (Doolittle and Surgenor 1962). However, one of the three repeats that occurs in tissue factor *inhibitor*

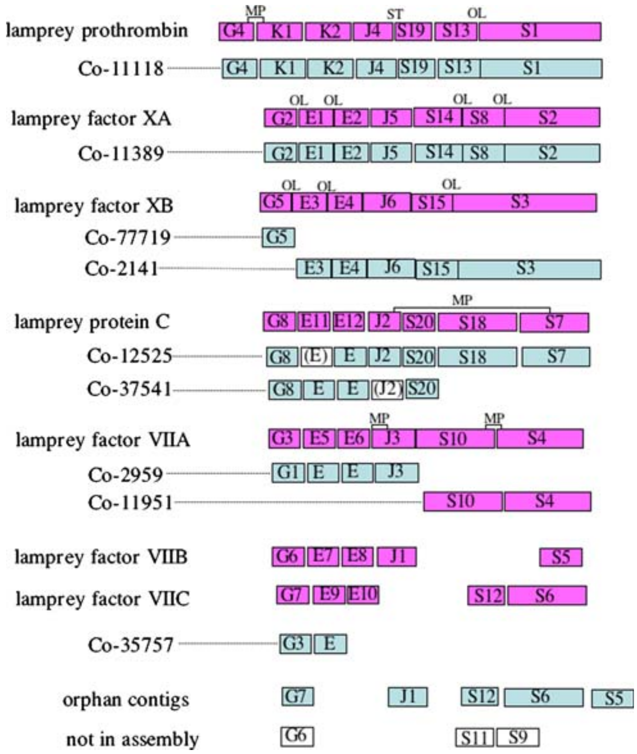

Fig. 7 Nine supercontigs were identified that contain hit-groups for vitamin K-dependent proteases in Trace database. The red blocks are the same as those in Fig. 5; light-blue blocks were found in supercontigs. Clear blocks either were not yet in the assembly or were in nonsequenced segments (NNNN-regions). Three of the supercontigs span entire genes from GLA domains to carboxyl domains (prothrombin, factor 10A, protein C). In two other cases, pairs of supercontigs cover the entire genes (factors VIIA and XB). Orphan contigs are those that contain only one hit-group. Three hit-groups are not yet included in the assembly

(TFI) was initially found in the Trace database (Trace ID 1470228595). The same sequence is found in the assembly along with the other two repeats on supercontig-12432. The sequence is about 45% identical to human TFI, which is actually a greater similarity than was found for the puffer fish-human comparison (Jiang and Doolittle 2003).

Positive identifications were also made for thrombomodulin, initially in the Trace database and then, more convincingly, in the draft assembly. In other vertebrates (puffer fish and mammals) thrombomodulin has an intron-free sequence, but it was unlikely that the full-length sequence (575 codons in the human version) would appear in a single sheared shotgun fragment in the Trace database. Nonetheless, one trace corresponded to a homologous amino-terminal lectin domain (Trace ID 1377464678) and two others encompassed identical strings of five EGF domains (Trace IDs 1382345176 and 1206195865). The full sequence, amounting to 700 residues, was found on supercontig-5373, with no introns, and included an appropriately positioned putative membrane-spanning segment. The sequence is only 30% identical to human thrombomodulin, not unreasonable considering that puffer fish thrombomodulin is only 37% identical to the human protein. A second lamprey thrombomodulin was found on supercontig-25172. In this case, the sequence

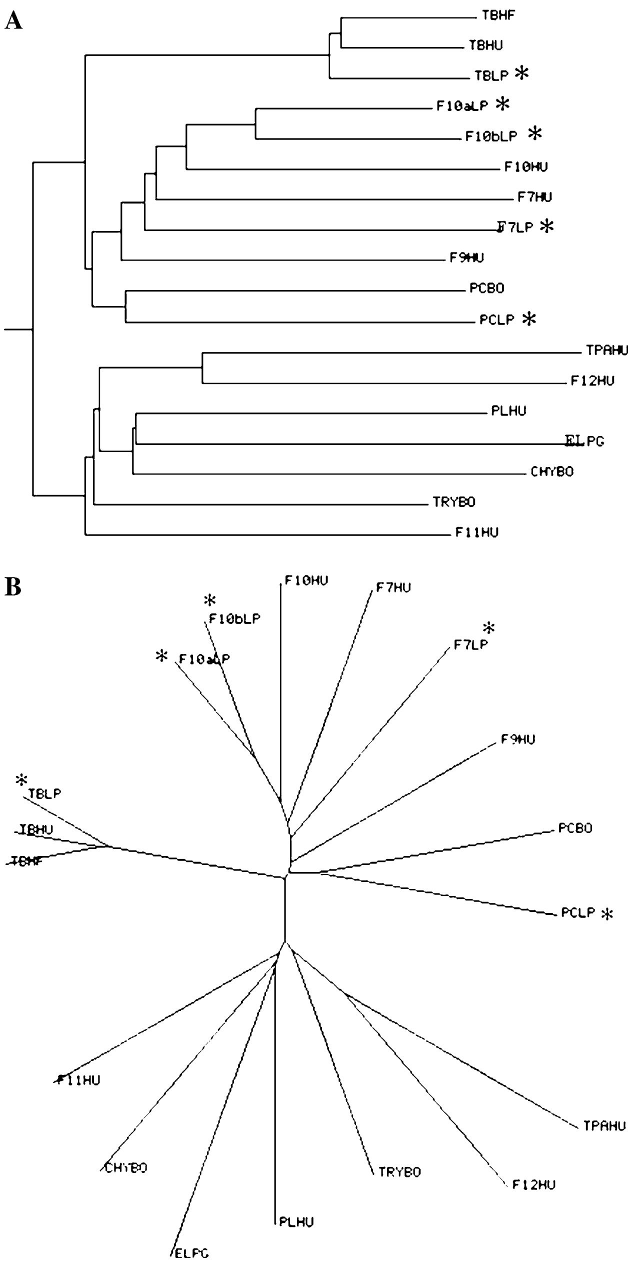

Fig. 8 Phylogenetic trees calculated from an alignment of complete serine protease regions of 18 widely assorted serine proteases, including putative sequences for lamprey prothrombin, protein C, two factors X, and a factor VII (lamprey sequences denoted with and asterisk). TRY, trypsin; CHY, chymotrypsin; F12, factor XII; TPA, tissue plasminogen activator; EL, elastase; F11, factor XI; PC, protein C; F9, factor IX; F7, factor VII; F10, factor X; TB, thrombin. HU, human, BO, bovine; PG, pig; HF, hagfish; LP, lamprey. (**A**) A tree generated by distance matrix method (Feng and Doolittle 1996). (**B**) An unrooted neighbor-joining tree based on the same alignment (Saitou and Nei 1987; Felsenstein 1989)

corresponding to the amino-terminal portion occurs in a nonsequenced region, and only 494 codons are present.

Discussion

The presence of genes in the lamprey genome that encode clotting factors was anticipated, a number of them having been isolated and/or cloned in the past. The basic events involving the thrombin-catalyzed conversion of fibrinogen to fibrin and its subsequent cross-linking and lysis were not an issue. The detailed pathway of thrombin generation was still undetermined, however, and whether or not all of the vitamin-K dependent factors are present in the jawless fish was not known.

We are now proposing that the lamprey has a reduced set of clotting factors, corresponding to what would have been in place before the duplication of two different kinds of protein, the vitamin K-dependent proteases, for one, and the factor V-VIII family, for the other. It is not yet possible to reconstruct the full sequence for the putative preduplication factor 5/8 gene. Five of the original 13 hit-groups were not used in the assembly, 2 were found linked to hephaestin-ceruloplasmin sequences, and 4 more are on orphan contigs. However, one—the very one that had been placed at the carboxyl-terminal end—was found to be linked to a discoid domain, a characteristic feature of factors V and VIII (Fig. 3). Importantly, none of the other hit-groups found in the assembly occur on contigs encoding discoid domains.

The evidence for there being only four genes in the hephaestin-ceruloplasmin-factor V–VIII family is based on there never being more than four different sequences for any set of aligned segments (Figs. 2 and 3). That only one of these is related to factors V and VIII is based on the majority of fragment sequences being more similar to hephaestin and ceruloplasmin and on intron locations coinciding with those of hephaestin and ceruloplamin in three of the sequences and with factors V and VIII in the fourth (Fig. 4). Additionally, the region occupied by a long stretch of low-complexity sequence in all known factors V and VIII is definitely not present in three of the reconstructed sequences. Finally, only one of the four reconstructed sequences was found to be associated with a discoidin sequence.

As for the unexpectedly small fraction of fragments that distinctly resemble factors V or VIII, it is possible that the more rapid rate of change of these proteins compared with hephaestin and ceruloplasmin contributes to the problem. It is well established that different paralogues can change at different rates in the wake of a gene duplication (Zhang et al. 2006). As such, when searching for paralogues in a genome that diverged not long after a gene duplication, a search of the faster-changing paralogue may find a better match with the slower-changing paralogue in the target organism (Fig. 9). This could explain why a search of human factor V and VIII sequences finds so many better matches with hephaestin and ceruloplasmin. Those difficulties notwithstanding, in the end the data are only consistent with a single copy of a gene corresponding to the precursor of clotting factors V and VIII.

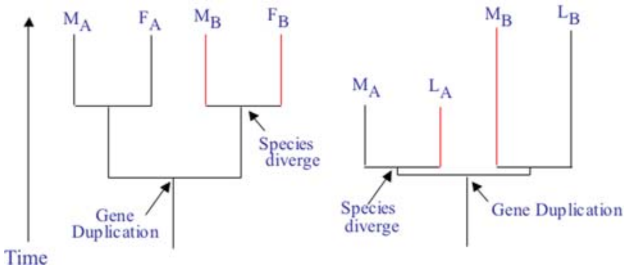

Fig. 9 Explanation of how searching with a fast-changing sequence can find a slower-changing paralogue rather than a faster-changing orthologue, particularly when the times of gene duplication and species divergence are near each other. The distance from M$_B$ to L$_A$ (paralogue) is shorter than the distance from M$_B$ to L$_B$ (orthologue)

Several kinds of evidence speak for the absence of the vitamin K-dependent protease factor IX. First, and most important, none of the candidate sequences gave factor IX as a top hit when subjected to reverse searching. Moreover, five of the six unique segments corresponding to the region in vitamin K-dependent proteases where cleavage activation occurs have a brace of cysteines universally observed in factors VII and X but which have not been found in factor IX in other vertebrates. On a more circumstantial level, of the four gene duplications that have previously been postulated as leading to the five vitamin K-dependent proteases, there is general agreement that the one giving rise to factors IX and X is the most recent (Doolittle and Feng, 1987; Doolittle 1993; Davidson et al. 2003a, b; Jiang and Doolittle 2003). In that same vein, the sequence similarities of factors IX and X are only slightly less than the resemblances observed between puffer fish and human sequences for those two factors (Jiang and Doolittle 2003), again suggesting that the duplication event occurred not long before the appearance of teleost fish. Although it is more difficult to prove the absence of a gene than its presence in a sequence database, we would cautiously propose that a gene corresponding to factor IX is not present in the lamprey genome.

The Matter of Whole-Genome Duplications

Although the notion that the vertebrate blood coagulation pathway is the result of a series of gene duplications is longstanding (Doolittle 1961; Doolittle and Feng 1987), more recently it has been suggested that the gene duplications responsible for some clotting factors may be linked to whole-genome duplication events (Davidson et al. 2003a, b). The possibility of polyploidy leading to whole-genome duplications was first raised by Ohno (1970), and it has been much discussed and hotly debated ever since (Sidow 1996; Smith et al. 1999; Hughes et al. 2001; Panopoulou et al. 2003; Dehal and Boore 2005; *inter alia*). Examples of preduplication genes in lamprey or hagfish—i.e., genes that are duplicated in other vertebrates—are well known, beginning with the observation that lampreys have a single-chained hemoglobin (Ingram 1963). More recent reports include high-mobility-group proteins (Sharman et al. 1997) and thyroid hormone and retinoid X rcceptors (Escriva et al. 2002).

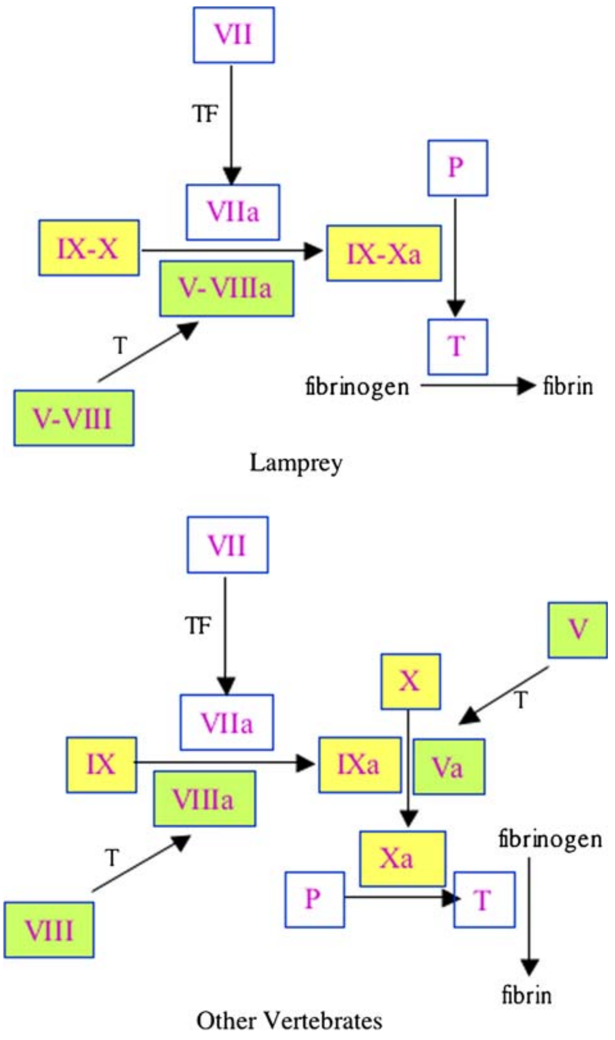

Fig. 10 Proposed scheme showing how genome doubling could give rise to simultaneous appearance of new orthologues for the ancestors of factors V/VIII and IX/X. In lampreys (top) there is only a single member of the factor V/VIII type (green) cofunctioning with a single vitamin K-dependent protease (yellow). The situation in other vertebrates (bottom) shows the result of gene doubling for both kinds of factor

In the case of the clotting scheme, a whole-genome duplication on the lineage leading to other vertebrates could advantageously result in the simultaneous production of additional copies of genes of the factors V–VIII type and the vitamin K-dependent type, as depicted in Fig. 10. It may only be a coincidence that the number of GLA domains has gone from 4 in the protochordate *Ciona intestinales* to 8 in the lamprey to 16 in the puffer fish, but the geometric progression is noteworthy.

In summary, the genomic picture presented here suggests that lampreys have a simpler clotting scheme than later diverging vertebrates. In particular, they appear to lack the equivalents of factors VIII (or V) and IX, suggesting that the gene duplications leading to these coagulation factors, synchronous or not, occurred after their divergence from other vertebrates.

In the end, the draft assembly of the lamprey genome, taken together with other sequences found in the Trace database, supports the suggestion that the lamprey genome lacks the separate equivalents of factors V and VIII (i.e., it has a preduplication gene) and a factor IX (i.e, it has a preduplication gene corresponding to factor X). A genuine, fully assembled lamprey genome should settle the matter unequivocally. Meanwhile, given the specific nature of our proposal, it may be possible to conduct biochemical experiments on lamprey blood that could address the question directly.

Acknowledgments This work was supported in part by National Institutes of Health Grant HL-81553. We thank Steve Culbertson for assistance in identifying mate-pairs. We are also are grateful to Jonathan Gitlin (Washington University, St. Louis) for sharing his unpublished findings on lamprey ceruloplasmin and to Steve Sommer (City of Hope) for unpublished cDNA sequences of two vitamin K-dependent proteases from lamprey.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Banfield DK, MacGillvray RTA (1992) Partial characterization of vertebrate prothrombin cDNAs:amplification and sequence analysis of the B chain of thrombin from nine different species. Proc Natl Acad Sci USA 89:2779–2783

Bohonus VL, Doolittle RF, Pontes M, Strong DD (1986) Complementary DNA sequence of lamprey fibrinogen β chain. Biochemistry 25:6512–6516

Carroll RL (1988) Vertebrate paleontology and evolution. W. H. Freeman, New York

Cripe LD, Moore KD, Kane WH (1992) Structure of the gene for human coagulation factor V. Biochemistry 31:3777–3785

Daimon M, Yamatani K, Igarashi M, Fukase N, Kawanami T, Kato T, Tominaga M, Sasaki H (1995) Fine structure of the human ceruloplasmin gene. Biochem Biophys Res Commun 208:1028–1035

Davidson CJ, Hirt RP, Lal K, Elgar G, Tuddenham EGD, McVey JH (2003a) Molecular evolution of the vertebrate blood coagulation network. J Thromb Haemost 1:1487–1494

Davidson CJ, Tuddenham EG, McVey JH (2003b) 450 Million years of hemostasis. Thromb Haemost 89:420–428

Dehal P, Boore JL (2005) Two rounds of whole genome duplication in the ancestral vertebrate. PLOS Biol 3:e314

Doolittle RF (1961) The comparative biochemistry of blood coagulation. Ph.D. dissertation. Harvard University

Doolittle RF (1987) Of Urfs and Orfs: a primer on how to analyze derived amino acid sequences. University Science Press, Mill Valley, CA

Doolittle RF (1993) The evolution of vertebrate blood coagulation: a case of Yin and Yang. Thromb Haemostasis 70:24–28

Doolittle RF, Feng DF (1987) Reconstructing the evolution of vertebrate blood coagulation from a consideration of the amino acid sequences of clotting proteins. Cold Spring Harbor Symp Quant Biol 52:869–874

Doolittle RF, Feng D-F (1990) Nearest neighbor procedure for relating progressively aligned amino acid sequences. Methods Enzymol 183:659–669

Doolittle RF, Surgenor DM (1962) Blood coagulation in fish. Am J Physiol 203:964–970

Doolittle RF, Oncley JL, Surgenor DM (1962) Species differences in the interaction of thrombin and fibrinogen. J Biol Chem 237:3123–3127

Escriva H, Manzon L, Youson J, Laudet V (2002) Analysis of lamprey and hagfish genes reveals a complex history of gene duplications during early vertebrate evolution. Mol Biol Evol 19:1440–1450

Felsenstein J (1989) PHYLIP—Phylogeny Inference Package (version 3.2). Cladistics 5:164–166

Feng D-F, Doolittle RF (1996) Progressive alignment and phylogenetic tree construction of protein sequences. Methods Enzymol 266:368–372

Gregory TR (2005) Animal Genome Size Database. Available at: http://www.genomesize.com

Hanumanthaiah R, Day K, Jagadeeswaran P (2002) Comprehensive analysis of blood coagulation pathways in Teleostei: evolution of coagulation factor genes and identification in zebrafish factor VIIi. Blood Cells Molecules Dis 29:57–68

Hellman NE, Gitlin JD (2002) Ceruloplasmin metabolism and function. Annu Rev Nutr 22:439–458

Henikoff JG, Henikoff S (1996) Blocks database and its applications. Methods Enzymol 266:88–105

Hughes AL, da Silva J, Friedman R (2001) Ancient genome duplications did not structure the human *Hox*-bearing chromosomes. Genome Res 11:771–780

Ichinose A, Takeya H, Esping E, Iwanaga S, Kisiel W, Davie EW (1990) Amino acid sequence of human protein Z, a vitamin K-dependent plasma glycoprotein. Biochem Biophys Res Commun 173:1139–1144

Ingram VM (1963) The hemoglobins in genetics and evolution. Columbia University Press, New York

Jagadeeswaran P, Gregory M, Zhou Y, Zon L, Padmanabhan K, Hanumanthaiah R (2000) Characterization of zebrafish full-length *prothrombin* cDNA and linkage group mapping. Blood Cells Molecules Dis 26:479–489

Jiang Y, Doolittle RF (2003) The evolution of vertebrate blood coagulation as viewed from a comparison of puffer fish and sea squirt genomes. Proc Natl Acad Sci USA 100:7527–7532

Kulman JD, Harris JE, Haldeman BA, Davie EW (1997) Primary structure and tissue distribution of two novel proline-rich γ-carboxyglutamic acid proteins. Proc Natl Acad Sci USA 94:9058–9062

Kulman JD, Harris JE, Nakazawa N, Ogasawara M, Satake M, Davie EW (2006) Vitamin K-dependent proteins in *Ciona intestinalis*, a basal chordate lacking a blood coagulation cascade. Proc Natl Acad Sci USA 103:15794–15799

Laize V, Martel P, Viegas CSB, Price PA, Cancela ML (2005) Evolution of matrix and bone gamma-carboxyglutamic acid proteins in vertebrates. J Biol Chem 280:26659–26668

Manfioletti G, Brancolini C, Avanzi G, Schneider C (1993) The protein encoded by a growth arrest-specific gene (gas6) is a new member of the vitamin K-dependent proteins related to protein S, a negative co-regulator in the blood coagulation cascade. Mol Cell Biol 13:4976–4985

Ohno S (1970) Evolution by gene duplication. Springer-Verlag, New York

Pan Y (1992) Studies on fibrinogen evolution. Ph.D. dissertation, University of California, San Diego

Pan Y, Doolittle RF (1992) cDNA sequence of a second fibrinogen α chain: an archetypal version alignable with full-length β and γ chains. Proc Natl Acad Sci USA 89:2066–2070

Panopoulou G, Hennig S, Groth D, Krause A, Poustka A, Herwig R, Vingron M, Lehrach H (2003) New evidence for genome-wide duplications at the origin of vertebrates using an amphioxus gene set and completed animal genomes. Genome Res 13:1056–1066

Price PA, Otsuka AA, Poser JW, Kristaponis J, Raman N (1976) Characterization of a gamma-carboxyglutamic acid-containing protein from bone. Proc Natl Acad Sci USA 73:1447–1451

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Sharman AC, Hay-Schmidt A, Holland PWH (1997) Cloning and analysis of an HMG gene from the lamprey *Lampetra fluviatilis*:gene duplication in vertebrate evolution. Gene 184:99–105

Sheehan J, Temple M, Gregory M, Hanumanthaiah R, Troyer D, Phan T, Thankavel B, Jagadeeswaran P (2001) Demonstration of the extrinsic coagulation pathway in teleostei: identification of zebrafish coagulation factor VII. Proc Natl Acad Sci USA 98:8768–8773

Sidow A (1996) Gen(om)e duplications in the evolution of early vertebrates. Curr Opin Gen Dev 6:715–722

Smith NGC, Knight R, Hurst LD (1999) Vertebrate genome evolution: A slow shuffle or a big bang?. BioEssays 21:697–703

Strong DD, Moore M, Cottrell BA, Bohonus VL, Pontes M, Evans B, Riley M, Doolittle RF (1985) Lamprey fibrinogen γ chain: cloning, cDNA sequencing and general characterization. Biochemistry 24:92–101

Syed BA, Beaumont NJ, Pate A, Maylor CE, Bayele HK, Joannu CL, Rowe PS, Evans RW, Srai SK (2002) Analysis of the human hephaestin gene and protein: comparative modelling of the N-terminus ecto-domain based upon ceruloplasmin. Prot Eng 15:205–214

Wang Y-Z, Patterson J. Gray JE, Yu C, Cottrell BA. Shimizu A, Graham D, Riley M, Doolittle RF (1989) Complete sequence of the lamprey fibrinogen α chain. Biochemistry 28:9801–9806

Zhang P, Gu Z, Li W-H (2006) Different evolutionary patterns between young duplicate genes in the human genome. Genome Biol 4:R56