THE EVOLUTIONARY RELATIONSHIPS OF THE ENZYMES INVOLVED IN BLOOD COAGULATION AND HEMOSTASIS *

David Hewett-Emmett

Department of Human Genetics
University of Michigan Medical School
Ann Arbor, Michigan 48109

John Czelusniak and Morris Goodman

Department of Anatomy
Wayne State University School of Medicine
Detroit, Michigan 48201

INTRODUCTION

Many of the blood coagulation factors circulate in the plasma as inactive precursors (zymogens), which upon activation by limited proteolysis become proteolytic enzymes with a high degree of substrate specificity.[1-3] Like the pancreatic digestive enzymes, the blood coagulation factors are serine proteases possessing an Asp, Ser, His at their active site.[1, 2] Unlike the pancreatic enzymes, most of the blood coagulation factors, upon activation, retain a large polypeptide chain linked by a disulfide-bridge to the chain containing the active site. Although crystals of thrombin have been grown,[4] no three-dimensional structure has yet been adduced, but it seems clear that the blood coagulation serine proteases will share many of the active-site features revealed by X-ray crystallography of the pancreatic enzymes, chymotrypsin A, trypsin, and elastase. Such studies have explained the difference in specificity of these enzymes [5, 6] however in their case almost all of the peptide bonds of the required specificity are cleaved; in the case of the blood coagulation factors only a few of the many arginyl bonds are cleaved.[1-3] In spite of the dearth of three-dimensional structures, amino acid sequence data are providing clues to understanding the diversity of this enzyme family.

Several investigators [7-9] have built gene phylogenies (genealogies) of the serine proteases using data sets of limited sizes. De Haën *et al.*[7] concluded that the presence of disulfide bridges and deletions/insertions might prove better phylogenetic markers than the sequences themselves. However, the recent publication of several new amino acid sequences provides us with an opportunity to shed some light on serine protease evolution by using the maximum parsimony method and extending earlier studies.[8, 9] The maximum parsimony approach assumes that evolution has taken the shortest course to reach the present array of diversity: it has been used effectively on such protein families as the hemoglobins,[10] intracellular calcium-binding proteins,[11] and carbonic anhydrase isozymes.[12] In this preliminary report, we describe the tree of lowest nucleotide replacement length that we found for each of the data sets used and draw some general conclusions about the evolution of the blood coagulation factors.

* *This work was supported* by National Science Foundation Grant DEB-7810717.

0077-8923/81/0370-0511 $01.75/0 © 1981, NYAS

Materials and Methods

The available amino acid sequences (Table 1) were grouped into four data sets and aligned to maximize homology, a procedure that even when computerized contains a large subjective input. To test whether aligned sequences showed significant homology with each other, an alignment statistic was used.[13] The aligned data sets created were:

1) SP—16 enzyme chain sequences which are more than 95% complete. For alignment, using the one-letter amino acid code (see Table 2). 2) SPIN-COMPL—The above 16 sequences with an additional 13 partial sequences. Since these align straightforwardly with the 16 complete sequences, the alignment for this data set is not illustrated, 3) VIT K—9 sequences (5 complete) representing the NH_2-terminal regions of the vitamin K-dependent blood coagulation factor zymogens. 4) KRINGLE—9 kringle loop structures of prothrombin and plasminogen aligned with 4 regions of putative homology in factor X, factor IX, protein C, and haptoglobin α chain.

For each data set, the maximum parsimony method was used to construct the tree that requires the fewest nucleotide replacements to explain the descent of extant sequences.[10, 11] Proteins are grouped so as to maximize the number of shared derived nucleotide replacements. Using a branch-swapping algorithm, many thousands of alternative trees are tested. No assumptions of constancy of rate of evolution are needed to construct genealogical trees by this method. The trees produced can be "rooted" subjectively (i.e. given a time dimension); however, inclusion of a bacterial serine protease enables the root to be placed on the branch to the eukaryotic serine proteases (see discussion [44]).

Results and Discussion

Tree Derived for 16 Serine Proteases (SP)

After testing many trees, that with the lowest nucleotide replacement length (1936 NR) is shown in Figure 1. The main features of the tree are:

1. The close grouping of the factors involved in fibrin-clot formation, supporting the view that blood coagulation was once a simple process involving perhaps a single thrombin-like enzyme that clotted a fibrinogen-like material.
2. The distant separation of plasmin from these blood coagulation factors. This was tested by submitting several alternative trees with widely different positions for plasmin. In all cases, plasmin was successively moved by the branch-swapping algorithm to its final position in the tree illustrated.
3. The finding that protein C was the first of the vitamin K-dependent factors to become a separate lineage, before the duplications that resulted in the factor IX and factor X lineages.

Haptoglobin is closely related to the blood coagulation factors. Like them, it is of hepatic origin, but during evolution it has lost the active-site His and Ser residues and its present role seems to be to bind hemoglobin virtually irreversibly.[9] Our tree is compatible with that of Kurosky *et al.*,[9] if they were to place the root of their tree in the same position as ours.

TABLE 1

AMINO ACID SEQUENCE DATA USED *

Protein	Species	Non-Enzyme Region	Enzyme Chain	References
Prothrombin	Human	322 (100%)	259 (100%)	14–16
	Ox	323 (100%)	259 (100%)	17–20
	Chicken	45 (14%)	0 (0%)	21
Factor X	Ox	191 (100%)	256 (100%)	22,23
Factor IX	Ox	181 (100%)	235 (100%)	24
Protein C	Ox	169 (100%)	242 (99%)	25
Factor VII	Ox	13 (?)	25 (~10%)	26
Protein S	Human	13 (?)	0 (0%)	27
	Ox	13 (?)	0 (0%)	27
Factor XI	Ox, Human	--------	36 (~14%)	28,29
Factor XII	Ox	--------	40 (~16%)	30
Plasminogen	Human	560 (100%)	230 (100%)	31,32
Haptoglobin	Human	83 (100%)	245 (100%)	9,33,34
	Rat	0 (0%)	40 (16%)	34
	Rabbit	0 (0%)	40 (16%)	34
	Dog	0 (0%)	40 (16%)	34
Kallikrein	Pig (pancreas)	--------	233 (96%)	35
Trypsinogen	Ox	--------	223 (100%)	cf. 7,36
	Pig	--------	223 (100%)	cf. 7
	Dogfish	--------	222 (100%)	cf. 7
Trypsinogen B	African lung fish	--------	147 (66%)	cf. 7
Cocoonase	Silkmoth	--------	30 (~12%)	37
RVV-V activator	Russell's viper	--------	14 (~6%)	38
Crotalase	Rattlesnake	--------	45 (~18%)	39
Complement CIr	Human	--------	20 (~8%)	40
Complement CIs	Human	--------	20 (~8%)	40
Complement Factor D	Human	--------	50 (~20%)	41
Group Specific Protease	Rat (intestine)	--------	224 (100%)	42
Chymotrypsinogen A	Ox	--------	230 (100%)	cf. 7,36
Chymotrypsinogen B	Ox	--------	230 (100%)	cf. 7,36
Proelastase B	Pig	--------	240 (100%)	cf. 7,36
Bacterial trypsin	*Streptomyces griseus*	--------	221 (100%)	43

* In the case of some zymogens, the non-enzyme chain region has been sequenced, but is not used in this presentation.

TABLE 2

ALIGNMENT OF DATA SET SP (16 SERINE PROTEASES, >95% COMPLETE) *

	16 20 22 27 30 32 40 42 50 57 60 70 80 90 100
HUMAN THROMBIN	IVEGSNAEIGMSPWQVMLFRKSPQ——ELLCGASLISNRWVLTAAHCLLYPPWNKNFTENDLLVRIGKHSRTRYERNIEKISMLEKIYIHPRYNWRENLD-
BOVINE THROMBIN	IVEGQDAEVGLSPWQVMLFRKSPQ——ELLCGASLISDRWVLTAAHCLLYPPWBKNFTVDDLLVRIGKHSRTRYERKVEKISMLDKIYIHPRYNWKENLD-
FACTOR XA	IVGGRDCAEGECPWQALLVNEEN———EGFCGGTILNEFYVLTAAHCLHQAKR———FT—————VRVG——DRNTQEGDEEMAHEVEMTVKHSRFV-KETYD-
FACTOR IXA	VVGGEDAERGQFPWQVLLHGEI————AAFCGGSIVNEKWVVTAAHCIKPGVK———IT—————VVAGEHNTEKPEPTEQKRN-VIRAIPYHSYNASIN——K
PROTEIN C	IVDGQEAGWGESPWQAVLLDSKK———KLVCGAVLIHVSWVLTVAHCXRKKLI——————————VRLGEYDMRRWESWEVDLD-IKEVIIHPNYTKSYSDN-
PLASMIN	VVGGCVAHPHSWPWQVSLRTRFG———MHFCGGTLISPEWVLTAAHCLEKSPRPSSYK—————VILGAHQEVNLEPHVQEIE-VSRLFLEP—————TRK——
HAPTOGLOBIN	ILGGHLDAKGSFPWQAKMVSHH————NLTTGATLINEQWLLTTAKNLFLNHSENAT——————AKDIAPTLTLYVGKKQLVE-IEKVVLHPNYSQV—————
OX TRYPSIN	IVGGYTCGANTVPYQVSLN——SG———YHFCGGSLINSQWVVSAAHCYKSGIQ——————————VRLGQDNINVVEGNQQFIS-ASKSIVHPSYNSNTLNN-
PIG TRYPSIN	IVGGYTCAANSIPYQVSLN——SG———SHFCGGSLINSQWVVSAAHCYKSRIQ——————————VRLGEHNIDVLEGNEQFIN-AAKIITHPNFNGNTLDN-
DOGFISH TRYPSIN	IVGGYECPKHAAPWTVSLN——VG———YHFCGGSLIAPGWVVSAAHCYQRRIQ——————————VRLGEHDISANEGDETYID-SSMVIRHPNYSGYDLDN-
CHYMOTRYPSIN A	IVNGEEAVPGSWPWQVSLQDKTG———FHFCGGSLINENWVVTAAHCGVTTSD—————————VVVAGEFDQGSSSEKIQKLK-IAKVFKNSKYNSLTINN-
CHYMOTRYPSIN B	IVNGEDAVPGSWPWQVSLQDSTG———FHFCGGSLISEDWVVTAAHCGVTTSD—————————VVVAGEFDQGLETEDTQVLK-IGKVFKNPKFSILTVRN-
ELASTASE	VVGGTEAQRNSWPSQISLQYRSGSSWAHTCGGTLIRQNWVMTAAHCVDRELT———FR—————VVVGEHNLNQNNGTEQYVG-VQKIVVHPYWNTDDVAA-
KALLIKREIN (GLAND.)	IIGGRECEKNSHPWQVAIYHYS————SFQCGGVLVNPKWVLTAAHCKNDNYE——————————VGWLRHNLFENENTAQFFG-VTADFPHPGFNLSADGKD
GROUP SP. PROTEASE	IIGGVESIPHSRPYMAHLDIVTEKGLRVICGGFLISRQFVLTAAHCKGREIT——————————VILGAHDVRKRESTQQKIK-VEKQIIHESYNSVPNL——
BACTERIAL TRYPSIN	VVGGTRAAQGEFPFMVRL——SMG——————CGGALYAQDIVLTAAHCVSGSGN———NT——————SITATGGVVDLQSAVKVLRSTKVLQAPGYNGTGK———

	102 110 120 122 127 130 136 140 150 157 160 168 170 180 180G
Human thrombin	—RDIALMKLKKPVAFSDYIHPVCLPNRETAAS-LLGAGYKGRVTGYGNLKSTVTADVGKGQPSVLQVVNLALVQRPVCKDS-TRI-RITDNM———
Bovine thrombin	—RDIALLKLKRPIELSDYIHPVCLPDKQTAAK-LLHAGFKGRVTGWGNRRETWTTSVAEVQPSVLQVVNLPLVERPVCKAS-TRI-RITNDM———
Factor Xa	—FDIAVLRLKTPIRF-RNVAPACLPEKDWAAE-TLQTKT-GIVSGFGR———TH-EKGRLSSTLKMLEVPYVDRSTCKLS-SSF-TITPNM———
Factor IXa	YSHDIALLELDEPLELNSYVTPICIADRDY————TNIF—SKFGYGYVSGWGKVFNRGRSASILQYLKVPLVDRATCLRS-TKF-SIYSHM———
Protein C	——DIALLRLAKPATLSQTIVPICLPDSGLSERKLTQVGQETVVTGWGYRDE———TKRNRTFVLSFIKVPVVPYXACVHA-MEN-KISENM———
Plasmin	——DIALLKLSSPAVITDKVIPACLPSPNY———VVADRTECFITGWGE———TQ—GTFGAGLLKEAQLPVIENKVCNRYEFLNGRVQSTE———
Haptoglobin	——DIGLIKLKQKVSVNERVMPICLPSKDYA————EVGRVGYVSGWGR———NA—NFKFTDHLKYVMLPVADQDQCIRH-YEGSTVPEKKTPKSPVG
Ox trypsin	——DIMLIKLKSAASLNSRVASISLPTSCA————SAGTQCLISGWGN———TKSSGTSYPDVLKCLKAPILSNSSCKSA-YPG-QITSNM———
Pig trypsin	——DIMLIKLSSPATLNSRVATVSLPRSCA————AAGTECLISGWGN———TKSSGSSYPSLLQCLKAPVLSDSSCKSS-YPG-QITGNM———
Dogfish trypsin	——DIMLIKLSKPAALNRNVDLISLPTGCA————YAGEMCLISGWGN———TM-DGAVSGDQLQCLDAPVLSDAECKGA-YPG-MITNNM———
Chymotrypsin A	——DITLLKLSTAASFSQTVSAVCLPSASD———DFAAGTTCVTTGWGL———TRYTNANTPDRLQQASLPLLSNTNCKK—YWGTKIKDAM———
Chymotrypsin B	——DITLLKLATPAQFSETVSAVCLPSADE———DFPAGMLCATTGWGK———TKYNALKTPDKLQQATLPIVSNTDCRK—YWGSRVTDVM———
Elastase	-GYDIALLRLAQSVTLNSYVQLGVLPRAGT———ILANNSPCYITGWGL———TR-TNGQLAQTLQQAYLPTVDYAICSSSSYWGSTVKNSM———
Kallikrein (gland.)	YSHDLMLLRLQSPAKITDAVKVLELPTQEP————ELGSTCEASGWGSI—EPGPDDFEFPDEIQCVQLTLLQNTFCAHA-BPB-KVTESM———
Group Sp. protease	—HDIMLLKLEKKVELTPAVNVVPLPSPSD———FIHPGAMCWAAGWGK———TG-VRDPTSYTLREVELRIMDEKACVDYRYYEYKF—Q———
Bacterial trypsin	——DWALIKLAQPIN———QP-TLKIATT——TAYNQGT-FTVAGWGA———NR-EGGSQQRYLLKANVPFVSDAACRSA-YGNELVANEE———

Table 2 (*continued*)

	182 189 191 195 200 201 210 214 216 220 226 232 240 245
Human thrombin	————FCAGYKPDEGKRGDACEGDSGGPFVMKSPFNNRWYQMGIVSWGE—GCDRDGKYGFYTHVFRLKKWI-QKVIDQFGE
Bovine thrombin	————FCAGYKPGEGKRGDACEGDSGGPFVMKSPYNNRWYQMGIVSWGE—GCDRNGKYGFYTHVFRLKKWI-QKVIDRLGS
Factor Xa	————FCAGY—DTQPE-DACQGDSGGPHV—TRFKDTYFVTGIVSWGE—GCARKGKFGVYTKVSNFLKWI-DKIMKARAGAAGSRGHSEAPATW
Factor IXa	————FCAGY—HEGGK-DSCQGDSGGPHV—TEVEGTSFLTGIISWGE—ECAMKGKYGIYTKVSRYVNWIKEKTKLT-
Protein C	————LCAGI—LGDPR-DACEGDSGGPMV—TFFRGTHFLVGLVSWGE—GCGRLYNYGVYTKVSRYLDWIYGHIKAQEAPLESQVP
Plasmin	————LCAGH—LAGGT-DSCQGDSGGPLV—CFEKDKQILQGVTSWGL—GCARPNKPGVYVRVSRFVTWI-EGVMRNN
Haptoglobin	VQPILNEHTFCAGM—SKYQE-DTCYGDAGSAFAVHDLEENTWYATGILSFDK——CSAVAEYGVYVKVTSIQNWV-QKTIAEN
Ox trypsin	————FCAGY—LEGGK-DSCQGDSGGPVV—CSGK——LQGIVSWGS—GCAQKNKPGVYTKVCNYVSWI-KQTIASN
Pig trypsin	————ICVGF—LEGGK-DSCQGDSGGPVV—CNGQ——LQGIVSWGY—GCAQKNKPGVYTKVCNYVNWI-QQTIAAN
Dogfish trypsin	————MCVGY—MEGGK-DSCQGDSGGPVV—CNGM——LQGIVSWGY—GCAERDHPGVYTRVCHYVSWI-HETIASV
Chymotrypsin A	————ICAG——ASGV—SSCMGDSGGPLV—CKKNGAWTLVGIVSWGSS-TCS-TSTPGVYARVTALVNWV-QQTLAAN
Chymotrypsin B	————ICAG——ASGV—SSCMGDSGGPLV—CQKNGAWTLAGIVSWGSS-TCS-TSTPAVYARVTALMPWV-QETLAAN
Elastase	————VCAG——GNGVR-SGCQGDSGGPLH—CLVNGQYAVHGVTSFVSRLGCNVTRKPTVFTRVSAYISWI-NNVIASN
Kallikrein (gland.)	————LCAGY—LPGGK-DTCMGDSGGPLI—CNGM——WQGITSWGHT-PCGSANKPSIYTKLIFYLDWI-BBTITENP
Group Sp. protease	————VCVGS—PTTLR-AAFMGDSGGPLL—CAGV——AHGIVSYGH—PDAKP—PAIFTRVSTYVPTI-NAVIN—
Bacterial trypsin	————ICAGY-PDTGGV-DTCQGDSGGPMFRKDNADE-WIQVGIVSWGY—GCARPGYPGVYTEVSTFASAI-ASAARTL

* Numbering based on chymotrypsin A. X=residue not identified.
—— = residue/region deleted.

Tree Derived for 29 Serine Proteases (SPINCOMPL)

After testing many trees, that with lowest nucleotide replacement length (2076 NR) is shown in FIGURE 2. The main features are:

1. The addition of 13 partial sequences has altered the arrangement of the pancreatic enzymes (trypsin, chymotrypsin, elastase, kallikrein) and the relationship of plasmin. At an intermediate stage of the work, before complement factor D and crotalase were added to the data set, plasmin still represented the earliest ancestral eukaryotic branch. It may well be that the addition of a complete complement factor sequence will be necessary to resolve the true position of plasmin in the genealogy. In general, experience with the hemoglobins has shown that additional sequences iron out discordances.[10]

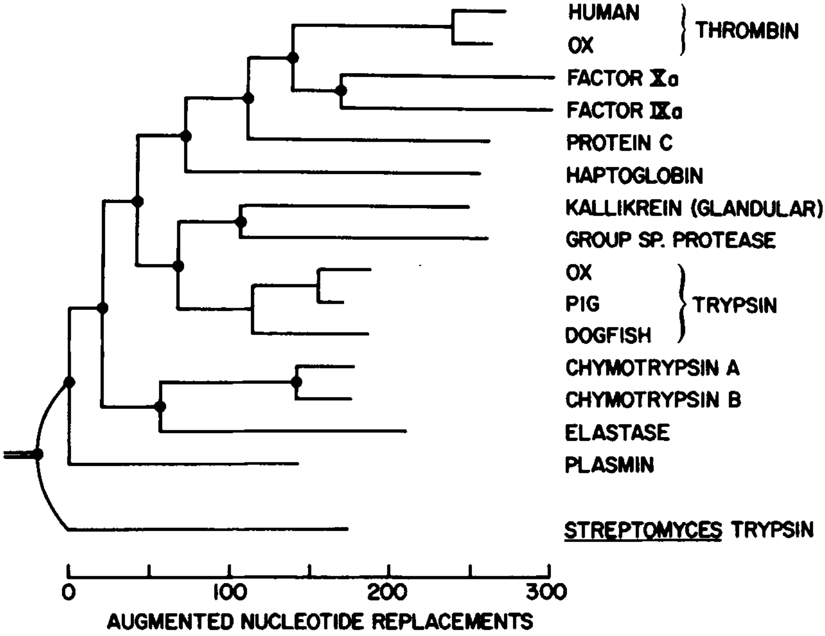

FIGURE 1. Genealogy of serine proteases based on data set SP (16 sequences more than 95% complete: TABLE 2). This tree has a nucleotide replacement length of 1936, the lowest found for this data set. The tree was "rooted" by use of the bacterial (*Streptomyces griseus*) trypsin. It should be noted that Hartley[45, 46] continues to believe that this gene is not of bacterial origin, having been inserted into the bacterial genome. It still seems to be the most distantly related of all the serine proteases examined in this data set. Other bacterial serine proteases, e.g. *Streptomyces griseus* protease B,[47] are clearly homologous to these proteases but many more insertions and deletions are required to align them. In earlier work (R. A. Marlar and D. Hewett-Emmett, 1976, unpublished), we included protease B and found that the "rooting" of the tree was identical, although plasmin was not available for inclusion at that time. Branch lengths are augmented to compensate for undetected multiple mutations in long separate lineages.[11] The branches are drawn to scale, their lengths being the augmented nucleotide replacements per enzyme chain. ● represent gene duplications.

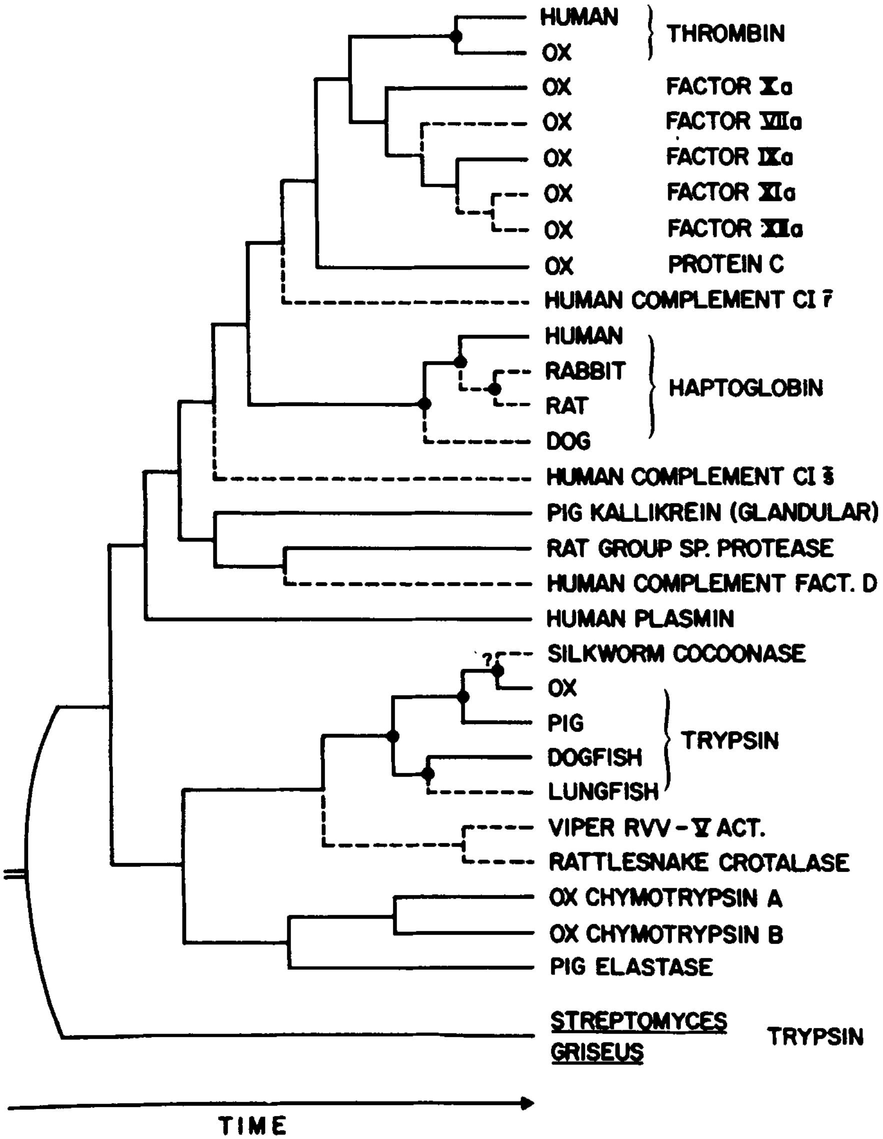

FIGURE 2. Genealogy of serine protease based on data set SPINCOMPL. This tree has a nucleotide replacement length of 2076, the lowest found for this data set. As in FIGURE 1, the tree was rooted by use of *S. griseus* trypsin. Broken lines indicate the 13 partial sequences whose positions in the tree are necessarily much less reliable. ● represent species divergencies; remaining bifurcations in the tree are gene duplications.

2. From the small stretches of sequence available, clotting factors VII, XI and XII show most affinity for factor IX. However, as with the complement factors, complete sequences will be needed to decide whether or not this is true. Factor XIIa probably possesses a disulfide bridge (residues 136–202) not present in the vitamin K-dependent coagulation factors.[30]

3. The snake venom proteases are clearly trypsin-like and not thrombin-like in their evolutionary relationship. This is supported by the probable presence of a disulfide bridge (residues 22–157 in TABLE 2) that is absent in all the hepatic serine proteases.[38, 39]

4. From the available data, it seems that the complement factors may not form a single grouping. It has been pointed out that factor D shows most similarity to the pancreatic serine proteases,[41] and in our tree it clusters with group-specific protease and kallikrein.

5. The relationship of haptoglobin to the blood coagulation factors is unaltered by the addition of the 13 partial sequences.

Tree Derived for NH_2-Terminal Regions of Vitamin K-Dependent Factors (VIT K)

After testing many trees, two with equal nucleotide replacement lengths (112 NR) were found. One tree split up the two partial protein S sequences and so the other tree is considered more likely to represent the true genealogy and is illustrated in FIGURE 3. The main features are:

1. The branching pattern is similar to that in FIGURES 1 and 2. However, the limited factor VII data indicate that it diverged earlier than was indicated in FIGURE 2. Until the full sequence is known this discrepancy cannot be resolved.

2. Based on equally weak evidence, protein S seems to be most closely related to factor X. Its function is not presently known.

Tree Derived for Kringle and Homologous Regions (KRINGLE)

The tree illustrated in FIGURE 4 is that with lowest nucleotide replacement length (504 NR). In this case, we know the order of the kringle regions on the plasminogen gene. Fitch[48] has pointed out that some phylogenies are incompatible with simple unequal crossover events. We have illustrated a mechanism involving unequal crossover events and gene deletions (7 events) whereby the observed kringle order can be derived. We disagree with Fitch[48] inasmuch as he does not allow gene deletion events in his scheme. It is of interest that the trees proposed by Young *et al.*[8] and Kurosky *et al.*[9] require at least 10 extra nucleotide replacements, although they require only 3 gene duplication events instead of 7 unequal crossover events in the scheme we advocate. The main features of the tree are:

1. The plasminogen kringle loops are most closely related to prothrombin kringle 1. It has been noted previously that prothrombin kringle 1 has been more conserved than kringle 2 or the vitamin K-dependent Ca^{2+}-binding region of prothrombin during mammalian evolution.[49, 50] No function has yet been ascribed to this region of either prothrombin or plasminogen however.

2. The branching pattern differs from those of Figures 1–3 inasmuch as factors X and IX do not share a period of evolution with prothrombin, assuming of course that the root has been placed correctly.

3. The putative kringle loops of factor X, factor IX, protein C, and haptoglobin are only weakly homologous with those of prothrombin and plasminogen using the Moore and Goodman[13] test. It is notable that on a less permissive visual test used previously,[50] factor X showed no detectable homology with either prothrombin kringle. Clearly if they are truly homologous, the kringle loops of protein C, factor X, and factor IX have diverged considerably while retaining significant homology to each other.

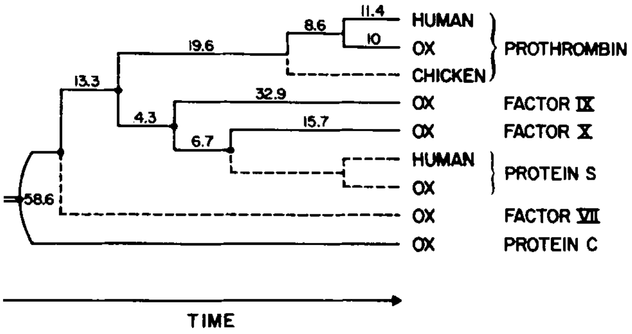

Figure 3. Genealogy of data set VIT K, representing regions homologous to residues 1–65 of bovine prothrombin (Table 3). This tree has a nucleotide replacement length of 112 and is one of two found with that length; the other split up the two protein S partial sequences and was considered less likely to represent the true genealogy. The tree was rooted using information derived from Figures 1 and 2, i.e. that protein C was the first separate lineage of those represented. This is supported by divergence data; protein C represents the longest branch. Broken lines indicate the 4 partial sequences. Branches have a time dimension and are *not* drawn to scale. The figures on the branches representing complete sequences are augmented nucleotide replacements per 100 codons. ● represent gene duplications.

Conclusions

This section will describe general conclusions based on the trees in Figures 1–3.

1. The order of gene duplication among the serine proteases involved in blood coagulation seems to be as follows:

Trypsin→*Trypsin–Plasmin*→*Trypsin–Thrombin*–Plasmin→
Trypsin–*Thrombin–Protein C*–Plasmin→
Trypsin–*Thrombin–Factor X*–Protein C–Plasmin→
Trypsin–Snake Venom Protease–Thrombin–*Factor X–Factor IX*–Protein C–Plasmin

The true place of factors VII, XI, XII, and the complement factors must await completion of their amino acid sequences.

2. The kringle loop structure regions provide strong evidence that plasminogen may be a hybrid gene, with kringle loops derived from prothrombin kringle 1 coding region having been fused to the plasmin light-chain coding region

TABLE 3

ALIGNMENT OF DATA SET VIT K, THE NH_2-TERMINAL REGIONS OF THE VITAMIN K-DEPENDENT BLOOD COAGULATION FACTOR *

		1 10 20 30 40 50 60 65
Human	Prothrombin	ANT-FLEE-VRKGNLERECVEETCSYEEAFEALESSTATDVFWAKYTA — CETARTPRDKLAACLE-GN
Ox	Prothrombin	ANKGFLEE-VRKGNLERECLEEPCSREEAFEALESLSATDAFWAKYTA — CESARNPREKLNECLE-GN
Chicken	Prothrombin	ANKGFLEE-MIKGNLERECLEETCNYEEAFEALESTVDTDAFWAKY
Ox	Factor X	ANS-FLEE-VKQGNLERECLEEACSLEEAREVFEDAEQTDEFWSKYKDGDQCEG ——— HPCLNQGH
Ox	Factor IX	YNSGKLEEFVR-GNLERECKEEKCSFEEAREVFENTEKTTEFWKQYVDGDQCES ——— NPCLNGGM
Ox	Protein C	ANS-FLEE-LRPGNVERECSEEVCEFEEAREIFQNTEDTMAFWSKYSDGEQCEDRPSGSPCDLPCCGRGK
Ox	Factor VII	AN-GFLEELL-PGSL
Human	Protein S	ANS-LLEE-XKQGNL
Ox	Protein S	ANT-LLEE-TKKGNL

* Numbering based on bovine profragment 1.
—— = residue/region deleted.

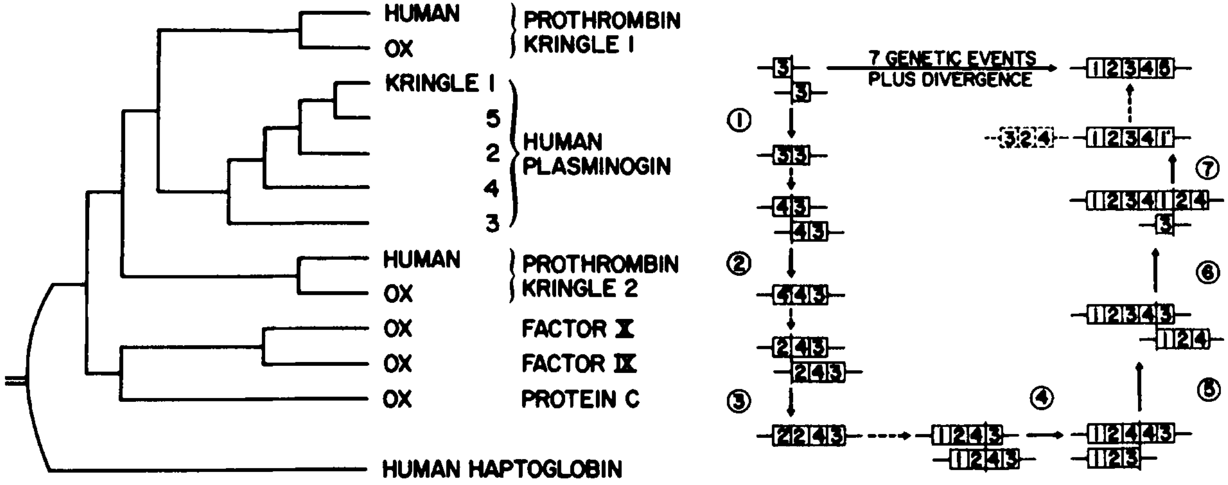

FIGURE 4. Hypothetical genealogy of the kringle structures of prothrombin and plasminogen and regions of putative homology in other serine protease zymogens (data set KRINGLE, TABLE 4). This tree has a nucleotide replacement length of 504, the lowest found for this data set. However at least 6 and probably 7 genetic events (unequal crossovers, etc.) are required to assemble the plasminogen kringles in the correct order on the chromosome. Alternative trees suggested elsewhere[8, 9] require at least 10 additional nucleotide replacements. The root of the tree is placed by assuming that haptoglobin represents the earliest ancestral branch. Since the homology of haptoglobin, factor X, factor IX, and protein C with the kringles of prothrombin and plasminogen is equivocal, this tree should be treated with caution with regard to genealogical conclusions. Branches are not drawn to scale although they have a time dimension.

Table 4

Alignment of "Kringle" Regions of Prothrombin and Plasminogen with Putative "Kringle" Regions of Other Serine Protease Zymogens *

		1 10 20 30 34A 40 50 60 70 79
Pro Kr 1	(H)	CAEGLGTNYRGNVSITRSGIECQLWRSRYPHKPE-INSTTHPGADLQENFCRNPDSSITGPWCYTTDPTA-RRQEC — STPV-C
	(B)	CAEGVGMNYRGNVSVTRSGIECQLWRSRYPHKPE-INSTTHPGADLRENFCRNPDGSITGPWCYTTSPTL-RREEC — SVPV-C
Pro Kr 2	(H)	CVPDRGQQYQGRLAVTTHGLPCLAWASAQAKALS-KHQDFNSAVQLVENFCRNPDGDEEGVWCYVAGKPG-DFGYC — DLNY-C
	(B)	CFPDRGREYRGRLAVTTSGSRCLAWSSEQAKALS-KDQDFNPAVPLAENFCRNPDGDEEGAWCYVADQPG-DFEYC — DLNY-C
Plas Kr 1		CKTGDGKNYRGTMSKTKNGITCQKWSSTSPHRPR-FSPATHPSEGLEENYCRNPDNDPQGPWCYTTDPEK-RYDYC — DILE-C
Kr 2		CMHCSGENYDGKISKTMSGLECQAWDSQSPHAHG-YIPSKFPNKNLKKNYCRNPDREL-RPWCFTTDPNK-RWELC — DIPR-C
Kr 3		CLKGTGENYRGNVAVTVSGHTCQHWSAQTPHTHN-TRPENFPCKNLDENYCRNPDGKR-APWCHTTNSQV-RWEYC — KIPS-C
Kr 4		CYHGDGQSYRGTSSTTTTGKKCQSWSSMTPHRHQ-KTPENYPNAGLTMNYCRNPDADK-GPWCFTTDPSV-RWEYC — NLKK-C
Kr 5		CMFGNGKGYRGKRATTVTGTPCQDWAAQEPHRHSIFTPETNPRAGLEKNYCRNPDGDVGGPWCYTTNPRK-LYDYC — DVPQ-C
Factor X		CKNGIG-DYTCTCAEGFEGKNCEFSTREI ———— CSLDNGGCDQFC-REERSEVRCSCAHGYVLGDDSKSCVS — TERFPC
Factor IX		CKTDIN-SYECWCQAGFEGTNCELDAT ———— CSIKNGRCKQFCKRDTDNKVVCSCTDGYRLAEDQKSCEP — AVPFPC
Protein C		CIHGLG-GFRCDCAEGWEGRFCLHEVRFS ———— NCSAEBGGCAHYC-MEEEGRRHCSCAPGYRLEDDHQLCVS — KVTFPC
Haptoglobin		VDSGNDVTDIADDGCPK-PPQIAHGYV-EHSVRYQC — KNYYKLR-TEGDGV — YTLNNEK-QWINKAVGDKLPE-C

* Numbering based on bovine prothrombin kringle 1.
— = residue/region deleted.

much later in time than the gene duplication generating the plasmin and thrombin lineages.

3. It seems clear that haptoglobin was once a hepatic serine protease that lost its proteolytic activity.[9] Our trees (FIGURES 1 and 2) are compatible with those of Kurosky *et al.*[9] provided that they alter the root of their tree. By contrast, the tree suggested by Doolittle[51] underestimates how closely related haptoglobin is to the hepatic serine proteases and, in particular, the blood coagulation factors.

ACKNOWLEDGMENTS

We appreciate the participation, at an earlier stage of the work, of Dr. Richard A. Marlar and the encouragement of Dr. Walter H. Seegers. We are particularly grateful to Dr. Alex Kurosky for providing us with a preprint of his work on haptoglobin and for discussion of his and our findings.

[**Note added in proof:** Since the meeting, several relevant amino acid sequences of serine proteases have been published. In particular, Bradshaw *et al.*[52] describe a chymotrypsin-like collagenase from the hepato-pancreas of the fiddler crab and a trypsin-like protease that comprises the γ-subunit of mouse nerve growth factor. Brunisholz *et al.*[53] provide substantial structural data on bovine and porcine plasminogen, which further emphasize the conservative nature of the kringle structure. Mole and Nieman[54] have partially sequenced human complement factor B, which is a novel serine protease and shows most similarity with plasminogen and rat intestine group-specific protease; as stated in the text, the addition of complement sequences may be necessary to identify the true phylogenetic relationships of plasminogen. Group-specific protease is now known to derive from atypical mast cells of the intestine and, in an excellent minireview on this topic, Woodbury and Neurath[55] describe a different but homologous protease from rat peritoneum and skeletal muscle mast cells. Interestingly, atypical mast cells do not contain heparin which is known to interact with lysine residues and the atypical mast cell protease contains almost 50% less lysine than the mast cell protease. Finally, Petersen *et al.*[56] have determined the amino-terminal sequence of protein Z from bovine plasma whose homology with the other vitamin K-dependent plasma clotting factors was previously inferred but not firmly proved by Prowse and Esnouf.[57] Protein Z contains γ-carboxyglutamic acid, but seems only distantly related to the other vitamin K-dependent factors.]

REFERENCES

1. SEEGERS, W. H., H. I. HASSOUNA, D. HEWETT-EMMETT, D. A. WALZ & T. J. ANDARY. 1975. Prothrombin and thrombin. Selected aspects of thrombin formation, properties, inhibition and immunology. Sem. Thromb. Hemost. **1:** 211–283.
2. DAVIE, E. W. & K. FUJIKAWA. 1975. Basic mechanisms in blood coagulation. Ann. Rev. Biochem. **44:** 799–829.
3. SUTTIE, J. W. & C. M. JACKSON. 1977. Prothrombin structure, activation and biosynthesis. Physiol. Rev. **57:** 1–70.
4. TSERNOGLOU, D., D. A. WALZ, L. E. McCOY & W. H. SEEGERS. 1974. An X-ray crystallographic study of thrombin. J. Biol. Chem. **249:** 999.

5. SEGAL, D. M., J. C. POWERS, G. H. COHEN, D. R. DAVIES & P. E. WILCOX. 1971. Substrate binding site in chymotrypsin A_γ. Biochemistry **10:** 3728–3738.
6. STROUD, R. M. 1974. A family of protein-cutting proteins. Sci. Am. **231:** 74–88.
7. DE HAËN, C., H. NEURATH & D. C. TELLER. 1975. The phylogeny of trypsin-related serine proteases and their zymogens. New methods for the investigation of distant evolutionary relationships. J. Mol. Biol. **92:** 225–259.
8. YOUNG, C. L., W. C. BARKER, C. M. TOMASELLI & M. O. DAYHOFF. 1978. Serine proteases. *In* Atlas of Protein Sequence and Structure. M. O. Dayhoff, Ed. Vol. 5 (Suppl. 3): 73–89. National Biomedical Research Foundation. Washington, D.C.
9. KUROSKY, A., D. R. BARNETT, T.-H. LEE, B. TOUCHSTONE, R. E. HAY, M. S. ARNOTT, B. H. BOWMAN & W. M. FITCH. 1980. Covalent structure of human haptoglobin: A serine protease homolog. Proc. Natl. Acad. Sci. USA **77:** 3388–3392.
10. GOODMAN, M., J. CZELUSNIAK, G. W. MOORE, A. E. ROMERO-HERRERA & G. MATSUDA. 1979. Fitting the gene linkage into its species linkage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Syst. Zool. **28:** 132–163.
11. GOODMAN, M., J.-F. PECHÈRE, J. HAIECH & J. G. DEMAILLE. 1979. Evolutionary diversification of structure and function in the family of intracellular calcium-binding proteins. J. Mol. Evol. **13:** 331–352.
12. TASHIAN, R. E., D. HEWETT-EMMETT & M. GOODMAN. Evolutionary diversity in the structure and activity of carbonic anhydrase. *In* Protides of the Biological Fluids. Colloquium No. 28. H. Peeters, Ed.: 153–156. Pergamon Press. Oxford.
13. MOORE, G. W. & M. GOODMAN. 1977. Alignment statistic for identifying related protein sequences. J. Mol. Evol. **9:** 121–130.
14. WALZ, D. A., D. HEWETT-EMMETT & W. H. SEEGERS. 1977. Amino acid sequence of human prothrombin fragments 1 and 2. Proc. Natl. Acad. Sci. USA **74:** 1969–1972.
15. BUTKOWSKI, R. J., J. ELION, M. R. DOWNING & K. G. MANN. 1977. Primary structure of human prethrombin 2 and α-thrombin. J. Biol. Chem. **252:** 4942–4957.
16. THOMPSON, A. R., D. L. ENFIELD, L. H. ERICSSON, M. E. LEGAZ & J. W. FENTON II. 1977. Human thrombin: Partial primary structure. Arch. Biochem. Biophys. **178:** 356–367.
17. REUTERBY, J., D. A. WALZ, L. E. MCCOY & W. H. SEEGERS. 1974. Amino acid sequence of O fragment of bovine prothrombin. Thromb. Res. **4:** 885–890.
18. HEWETT-EMMETT, D., L. E. MCCOY, H. I. HASSOUNA, J. REUTERBY, D. A. WALZ & W. H. SEEGERS. 1974. A partial gene duplication in the evolution of prothrombin? Thromb. Res. **5:** 421–430.
19. HEWETT-EMMETT, D., D. A. WALZ, J. REUTERBY, L. E. MCCOY & W. H. SEEGERS. 1975. The amino acid sequence of PR fragment (NH_2-terminal fragment) of bovine prothrombin. Thromb. Res. **7:** 227–234.
20. MAGNUSSON, S., T. E. PETERSEN, L. SOTTRUP-JENSEN & H. CLAEYS. 1975. Complete primary structure of prothrombin. *In* Proteases and Biological Control. E. Reich, D. B. Rifkin & E. Shaw, Eds.: 123–149. Cold Spring Harbor Laboratory. Cold Spring Harbor, N. Y.
21. WALZ, D. A. 1978. Comparative aspects of prothrombin activation. Bibliotheca Haemat. **44:** 8–14.
22. TITANI, K., K. FUJIKAWA, D. L. ENFIELD, L. H. ERICSSON, K. A. WALSH & H. NEURATH. 1975. Bovine factor X (Stuart factor): Amino acid sequence of heavy chain. Proc. Natl. Acad. Sci. USA **72:** 3082–3086.

23. Enfield, D. L., L. H. Ericsson, K. Fujikawa, K. A. Walsh, H. Neurath & K. Titani. 1979. Amino acid sequence of the light chain of bovine factor X, (Stuart factor). Biochemistry **19:** 659–667.

24. Katayama, K., L. H. Ericsson, D. L. Enfield, K. A. Walsh, H. Neurath, E. W. Davie & K. Titani. 1979. Comparison of amino acid sequence of bovine coagulation factor IX (Christmas factor) with that of other vitamin K-dependent plasma proteins. Proc. Natl. Acad. Sci. USA **76:** 4990–4994.

25. Fernlund, P. & J. Stenflo. 1980. Amino acid sequence of bovine protein C. *In* Vitamin K Metabolism and Vitamin K-dependent Proteins. J. W. Suttie, Ed.: 84–88. University Park Press. Baltimore, Md.

26. Kisiel, W., K. Fujikawa & E. W. Davie. 1977. Activation of bovine factor VII (proconvertin) by factor XIIa (activated Hageman factor). Biochemistry **16:** 4189–4194.

27. Discipio, R. G. & E. W. Davie. 1979. Characterization of protein S, a γ-carboxyglutamic acid containing protein from bovine and human plasma. Biochemistry **18:** 899–904.

28. Koide, T., M. A. Hermodson & E. W. Davie. 1977. Active site of bovine factor XI (plasma thromboplastin antecedent). Nature **266:** 729–730.

29. Kurachi, K. & E. W. Davie. 1977. Activation of human factor XI (plasma thromboplastin antecedent) by factor XIIa (activated Hageman factor). Biochemistry **16:** 5831–5839.

30. Fujikawa, K., K. Kurachi & E. W. Davie. 1977. Characterization of bovine factor XIIa (activated Hageman factor). Biochemistry **16:** 4182–4188.

31. Wiman, B. 1977. Primary structure of the β-chain of human plasmin. Eur. J. Biochem. **76:** 129–137.

32. Sottrup-Jensen, L., H. Claeys, M. Zajdel, T. E. Petersen & S. Magnusson. 1978. The primary structure of human plasminogen isolation of two lysine-binding fragments and one "mini"-plasminogen (MW, 38,000) by elastase-catalysed-specific limited proteolysis. *In* Progress in Chemical Fibrinolysis and Thrombolysis. J. F. Davidson, R. M. Rowan, M. M. Samama & P. C. Desnogens, Eds. Vol. 3: 191–209. Raven Press. New York.

33. Kurosky, A., D. R. Barnett, M. A. Rasco, T.-H. Lee & B. H. Bowman. 1974. Evidence of homology between the β-chain of human haptoglobin and the chymotrypsin family of serine proteases. Biochem. Genet. **11:** 279–293.

34. Kurosky, A., H.-H. Kim & B. Touchstone. 1976. Comparative sequence analysis of the N-terminal region of rat, rabbit and dog haptoglobin β-chains. Comp. Biochem. Physiol. **55:** 453–459.

35. Tschesche, H., G. Mair, G. Godec, F. Fiedler, W. Ehret, C. Hirschauer, M. Lemon & H. Fritz. 1979. The primary structure of porcine glandular kallikrein. Adv. Exp. Med. & Biol. **120:** 245–260.

36. Dayhoff, M. O. 1972. Proteases related to trypsin. *In* Atlas of Protein Sequence and Structure. M. O. Dayhoff, Ed. Vol. **5:** D99–D111. National Biomedical Research Foundation. Washington, D.C.

37. Kramer, K. J., R. L. Felsted & J. H. Law. 1973. Cocoonase. V. Structural studies on an insect serine protease. J. Biol. Chem. **248:** 3021–3028.

38. Kisiel, W. 1979. Molecular properties of the factor V-activating enzyme from Russell's viper venom. J. Biol. Chem. **254:** 12230–12234.

39. Baumgartner, R., T. Fletcher, I. Theodor, S. S. Bajwa, F. S. Markland & H. Pirkle. 1980. Amino acid sequences in crotolase, a thrombin-like enzyme from the venom of *Crotalus adamanteus*. Fed. Proc. **39:** 1027 (Abst #4001).

40. Sim, R. B., R. R. Porter, K. B. M. Reid & I. Gigli. 1977. The structure and enzymatic activities of the C1r and C1s subcomponents of C1, the first component of human serum complement. Biochem. J. **163:** 219–227.

41. Volanakis, J. E., A. S. Bhown, J. C. Bennett & J. E. Mole. 1980. Partial

amino acid sequence of human factor D: Homology with serine proteases. Proc. Natl. Acad. Sci. USA **77:** 1116–1119.

42. Woodbury, R. G., N. Katunuma, K. Kobayashi, K. Titani & H. Neurath. 1978. Covalent structure of a group-specific protease from rat small intestine. Biochemistry **17:** 811–819.
43. Olafson, R. W., L. Jurášek, M. R. Carpenter & L. B. Smillie. 1975. Amino acid sequence of *Streptomyces griseus* trypsin. Biochemistry **14:** 1168–1177.
44. Cook, C. N. & D. Hewett-Emmett. 1974. The uses of protein sequence data in systematics. *In* Prosimian Biology. R. D. Martin, G. A. Doyle & A. C. Walker, Eds.: 937–958. Duckworth. London.
45. Hartley, B. S. 1970. Homologies in serine proteinases. Phil. Trans. Royal Soc. London B **257:** 77–87.
46. Hartley, B. S. 1979. Evolution of enzyme structure. Proc. Royal Soc. London B **205:** 443–452.
47. Johnson, P. & L. B. Smillie. 1974. The amino acid sequence and predicted structure of *Streptomyces griseus* protease B. FEBS Lett. **47:** 1–6.
48. Fitch, W. M. 1977. Phylogenies constrained by the crossover process as illustrated by human hemoglobins and a thirteen-cycle, eleven-amino-acid repeat in human apolipoprotein A-1. Genetics **86:** 632–644.
49. Hewett-Emmett, D., D. A. Walz & W. H. Seegers. 1977. Evolutionary and functional observations on the primary structure of the non-thrombin region (residues 1–273) of human prothrombin. Biochem. Soc. Trans. (UK) **5:** 1452–1455.
50. Hewett-Emmett, D. 1978. Amino acid sequence homology and the vitamin K-dependent proteins. Bibliotheca Haemat. **44:** 94–104.
51. Doolittle, R. F. 1979. 3rd Edit. Protein evolution. *In* The Proteins. H. Neurath & R. L. Hill, Eds.: 1–118. Academic Press, Inc. New York.
52. Bradshaw, R. A., G. A. Grant, K. A. Thomas & A. Z. Eisen. 1980. Mouse NGF γ subunit and crab collagenase: Two serine proteases of unusual function. *In* Protides of the Biological Fluids. Colloquium No. 28. H. Peeters, Ed.: 119–122. Pergamon Press. Oxford.
53. Brunisholz, R., P. Moser, J. Schaller & E. Rickli. 1980. Partial sequence comparison between human, bovine and porcine plasminogen. *In* Protides of the Biological Fluids. Colloquium No. 28. H. Peeters, Ed.: 103–106. Pergamon Press. Oxford.
54. Mole, J. E. & M. Niemann. 1980. Structural evidence that complement factor B constitutes a novel class of serine protease. J. Biol. Chem. **255:** 8472–8476.
55. Woodbury, R. G. & H. Neurath. 1980. Structure, specificity and localization of the serine proteases of connective tissue. FEBS Lett. **114:** 189–196.
56. Petersen, T. E., H. C. Thøgersen, L. Sottrup-Jensen, S. Magnusson & H. Jornvall. 1980. Isolation and *N*-terminal sequence of protein Z, a γ-carboxyglutamic acid containing protein from bovine plasma. FEBS Lett. **114:** 278–282.
57. Prowse, C. W. & M. P. Esnouf. 1977. The isolation of a new warfarin-sensitive protein from bovine plasma. Biochem. Soc. Trans. (UK) **5:** 255–256.