PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Molecular Evolution of Serine Protease and Its Inhibitor with Special Reference to Domain Evolution

Takashi Gojobori and Kazuho Ikeo

Phil. Trans. R. Soc. Lond. B 1994 **344**, 411-415
doi: 10.1098/rstb.1994.0080

Molecular evolution of serine protease and its inhibitor with special reference to domain evolution

TAKASHI GOJOBORI AND KAZUHO IKEO

DNA Research Center, National Institute of Genetics, Mishima 411, Japan

SUMMARY

The evolution of serine protease and its inhibitor are discussed with special reference to domain evolution. It is now known that most proteins are composed of more than one functional domain. Because serine proteases such as urokinase and plasminogen are made of various functional domains, these proteins are typical examples of the so-called mosaic proteins. When Kringle domains in serine proteases and a Kunitz-type protease inhibitor domain in the amyloid β precursor protein in Alzheimer's disease patients were examined by the molecular evolutionary analysis, the phylogenetic trees constructed showed that these functional domains had undergone dynamic changes in the evolutionary process. In particular, these domains are evolutionarily movable. Thus, it is concluded that various functional domains evolved independently of each other and that they have been shuffled to create the existent mosaic proteins. This conclusion leads us to the reasonable speculation that those functional domains must have been minigenes possibly at the time of primordial life or the origin of life. We call these minigenes 'ancestral minigenes'. Every effort should be made to answer the question about the minimum set of ancestral minigenes that must have existed and must have been needed for maintaining life forms. The DNA sequence database is useful for making attempts to answer such difficult but significant questions.

1. INTRODUCTION

In general, a serine protease consists of not only a protease domain but also other functional domains. For example, uPA (urokinase) can be separated into two chains, A-chain and B-chain. Although the B-chain corresponds to a protease domain having serine at its active site, the A-chain contains various functional domains such as an EGF (epidermal growth factor)-like domain and the so-called Kringle domain (Pathy 1985). Kringle is a characteristic secondary structure of a protein molecule made of three pairs of cysteine–cysteine bonds. We may call this kind of protein as a 'mosaic protein', because it has more than one different domain.

On the other hand, a serine protease inhibitor is generally much smaller than serine proteases, and it usually has a single domain which functions as a protease inhibitor. Very recently, it was found that a serine protease inhibitor domain called the Kunitz-type inhibitor has been inserted into beta-amyloid precursor proteins (βAPP) in the brains of Alzheimer's disease patients (Ponte *et al.* 1991). It seems that the protease inhibitor has become one of the functional domains in βAPP during evolution.

It is of particular interest to study how such mosaic proteins evolved and what the evolutionary origin of the inserted protease inhibitor domain is. In the present paper, we shall discuss the evolution of serine proteases with special reference to Kringle domains. We shall also discuss the evolutionary relationship between the inserted protease inhibitor domain in βAPP and other protease inhibitors. We shall then show that these functional domains must have undergone remarkably dynamic evolution. Moreover, we point out that these functional domains are evolving independently of each other.

Retrospectively speaking, these functional domains may have been complete but smaller genes in the evolutionary past. We may call those ancestral genes 'ancestral minigenes'. In the world of primordial life, there may have existed at least a minimum set of ancestral minigenes that were necessary for sustaining life forms. During evolution, these minigenes must have been duplicated repeatedly and must have been combined with other minigenes, so that many mosaic proteins are found in the existent proteins. Along with accumulation of mutations such as nucleotide substitutions, the existent proteins must have survived until the present time.

Finally, we discuss the possibility that we can identify the minimum set of ancestral minigenes by surveying the DNA sequence database which contains a vast amount of nucleotide sequence data for all the organisms so far examined.

2. MOLECULAR EVOLUTION OF SERINE PROTEASE

(*a*) *Kringle domains in serine protease*

Let us focus our attention mainly on the serine proteases involved in the blood-coagulation system. A serine protease is composed of several functional

Phil. Trans. R. Soc. Lond. B (1994) **344,** 411–415
Printed in Great Britain

domains such as EGF domains, Kringle domains, and catalytic domains. The catalytic domains are also called 'protease domains'.

It is of particular interest to note that the number of Kringle domains varies with the proteins. In fact, uPA (urokinase) has a single Kringle domain, although tPA (tissue-type plasminogen activator) and prothrombin have respectively two Kringle domains. Plasminogen has five Kringles. The Kringle domain, which is about 80 amino acids long, was named after the shape of a northern European confectionery which is very similar to this characteristic secondary structure.

(*b*) *Kringle domains in other proteins*

Recently, four Kringles were discovered in HGF (hepatocyte growth factor) (Nakamura *et al.* 1989). Interestingly, HGF contains a protease domain as well as an EGF domain, although HGF does not function as a protease. It is also known that the HGF activator, which activates HGF, also contains a single Kringle (Miyazawa *et al.* 1993). Surprisingly enough, human apolipoprotein(a) which is a lipid transporting protein contains 38 Kringles (McLean *et al.* 1987). Although this protein also possesses a protease domain, it is experimentally known that the protease domain does not have functional activities.

The multiple alignment of amino acid sequences for Kringle domains from various organisms shows that six cysteine residues, which form three pairs of cysteine-cysteine bonds, are strongly conserved together with other some amino acids. A phylogenetic tree for Kringle domains constructed by the Neighbour-Joining (NJ) method (Saitou & Nei 1987) shows that the number of Kringles changed dynamically during evolution.

(*c*) *Evolutionary process of Kringle domains*

Figure 1 shows a summary of the evolutionary processes of Kringle domains which have been inferred from the phylogenetic tree (Ikeo *et al.* 1991). First, a single Kringle domain duplicated itself about 500 million years (Ma) ago. Then, the ancestral gene of plasminogen having five Kringles emerged from its ancestor approximately 300 Ma ago. About 80 Ma ago, the entire gene of ancestral plasminogen was duplicated into two genes, so that both genes contained five Kringles. One of them became the present plasminogen as it was, but the other underwent very dynamic evolution. In particular, the first three Kringle domains must have been deleted from the plasminogen-type ancestor protein. As a result, only the fourth and fifth Kringles remained in the duplicate. Then, the remaining fourth Kringle was duplicated repeatedly many times to become apolipoprotein(a). These repeated duplications seem to have taken place within only the last 5 Ma. These features of the evolutionary processes of Kringle domains were supported by a validity test such as the bootstrapping method of the phylogenetic tree.

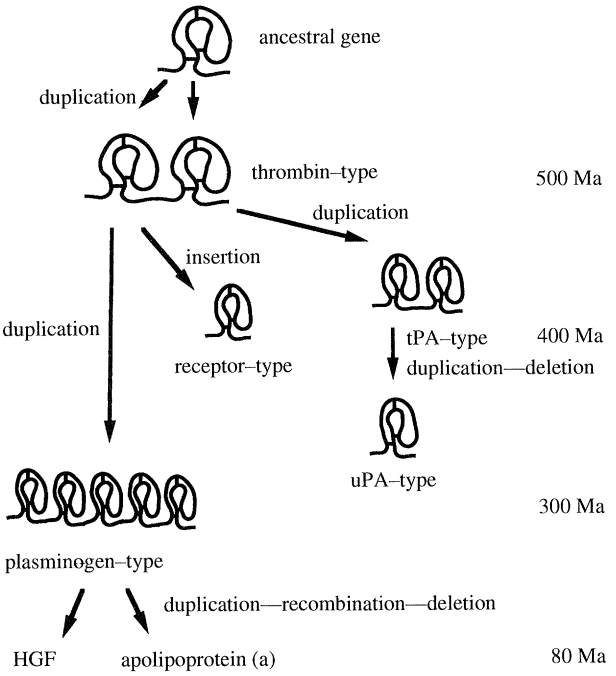

Figure 1. The evolutionary process of Kringle domains in serine proteases.

As mentioned above, the Kringle domain is evolving dynamically in terms of number. Moreover, the Kringle domain seems to be evolving as a unit, because incomplete Kringle domains such as a half of a Kringle domain and one fourth of the domain have never been observed. Of course, we cannot rule out the possibility that the incomplete Kringle domains were selected out quickly because of functional disadvantages. Because the serine protease has several domains as well as Kringle domains, such as EGF and protease domains, we constructed a phylogenetic tree for Kringle domains and protease domains separately by use of the same data set. Interestingly enough, two topologies for the Kringles and protease domains are totally different from each other. This strongly suggests that the Kringles and protease domains are evolving independently from each other.

(*d*) *Kringle domain as an evolutionarily movable unit*

Recently, a Kringle domain has been found in a receptor gene, *ror*, which may be involved in signal transduction in connection with the nervous system (Masiakowski & Carroll 1992). This gene does not have a protease domain at all. Moreover, its gene product is apparently a transmembrane-bound protein, but this gene is not a protease gene. Thus, the Kringle domain seems to have been inserted into this gene in the evolutionary process after the ancestral gene of *ror* had emerged. It means that the Kringle domains are evolutionarily movable.

From these observations about the evolutionary process of Kringle domains, it is clear that the functional domains are evolving dynamically. Russel Doolittle also proposed a similar hypothesis that the

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

domain is evolutionarily movable (Doolittle & Bork 1993). The present study supports this idea very strongly.

3. MOLECULAR EVOLUTION OF PROTEASE INHIBITORS

(*a*) *Protease inhibitor domain inserted into βAPP*

It is well known that the brain of an Alzheimer's disease patient contains big cavities and is accumulating specific proteins called 'βAPP' (β amyloid protein precursor) (Muller-Hill & Beyrenther 1989). Although there are two different sizes of βAPPs, APP695 and APP751, it has been said that a patient tends to produce APP751 at a larger quantity than APP695. APP751 has a molecular mass larger than APP695, because APP751 contains an additional domain, compared with APP695 (Kang *et al.* 1987; Ponte *et al.* 1988). It is known that both APP695 and APP751 are encoded by a single gene and that these two different proteins are expressed by the alternative splicing mechanism. In the human genome, this additional domain is encoded by a single exon which is located within the βAPP gene. Thus, this suggests that the exon encoding for the additional domain in APP751 was inserted into the APP695 gene in the evolutionary past.

This additional domain is now found to be a Kunitz-type protease inhibitor domain. In fact, the amino acid sequence of the inserted domain in APP751 can inhibit the activity of a serine protease, and it has a characteristic secondary structure called the 'Kunitz-type' or 'mini-kringle structure'. The same type of structures can be seen in BPTI (bovine pancreatic trypsin inhibitor), snake toxins, LACI (lipoprotein associated coagulation inhibitor) (Wun *et al.* 1988), and so forth (Ikeo *et al.* 1992).

(*b*) *Biological function of the inserted domain*

Let us ask the following questions. What is the biological function of the Kunitz-type protease inhibitor domain that was inserted into βAPP? What is the evolutionary origin of the inserted domain? When was the protease inhibitor domain of APP751 inserted into the ancestral form of APP695? To attempt to answer these interesting questions, a phylogenetic tree was constructed by use of almost all amino acid sequences presently available for Kunitz-type protease inhibitors (Ikeo *et al.* 1992).

The phylogenetic tree obtained shows that the inserted domain of APP751 has the closest common ancestor with trypstatin and IαTI (inter α trypsin inhibitor). Trypstatin inhibits the activity of tryptase, a novel membrane-bound serine protease in human T4 lymphocytes. In fact, the V3 domain of HIV-1 (human immunodeficiency virus type 1) envelope glycoprotein (gp120) is able to bind tryptase specifically. This binding activity was selectively blocked by trypstatin (Kido *et al.* 1991). Although a biological function of the inserted domain has not been known experimentally, the phylogenetic tree suggests that the inserted domain may have a function as an inhibitor of a membrane-bound serine protease in the cells related to the signal transduction in the immune system or possibly the self-immune mechanisms (Ikeo *et al.* 1992).

(*c*) *Evolutionary process of the Kunitz-type protease inhibitor domains*

It is also shown that the inserted domain of βAPP has a long evolutionary history. As shown in figure 2, the evolutionary process of the Kunitz-type protease inhibitors can be summarized in the following. The Kunitz-type protease inhibitor domain seems to have diverged from its common ancestor about 450 Ma ago. Then, the domain duplication must have taken place, and one of the duplicated domains was inserted into βAPP about 270 Ma ago. Because the divergence time of 270 Ma corresponds roughly to the time of species divergence between human and chicken, it follows that human βAPP must have retained this inserted domain for a long time.

As discussed above, the Kunitz-type protease inhibitor domain has also undergone dynamic changes in the evolutionary process. Moreover, this domain also appears to be evolutionarily movable because it was inserted into βAPP several hundred years ago.

4. A MINIMUM SET OF ANCESTRAL MINIGENES

(*a*) *The origin and evolution of functional domains*

We have observed that the Kringle domains and the Kunitz-type protease inhibitor domains have evolved

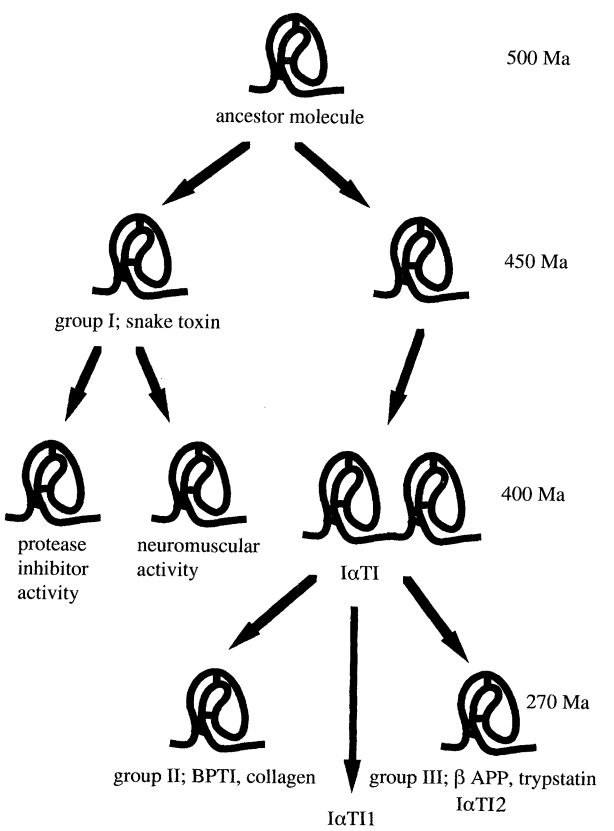

Figure 2. The evolutionary process of Kunitz-type protease inhibitors with special reference to the domain inserted into βAPP in Alzheimer's disease patients.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

independently and dynamically in their respective evolutionary processes. These observations lead us to the following interesting speculations (figure 3). In the evolutionary past, each domain could have been a minigene which had been in charge of a particular biological function. These minigenes must have a common ancestor called 'ancestral minigene'. By duplication, recombination, and possibly other evolutionary mechanisms at the molecular level, different minigenes must have been combined with each other and must have been duplicated many times, resulting in creation of the present proteins as mosaic proteins. Of course, mutations such as nucleotide substitutions must have accumulated. In many cases, these mutations must have made a substantial contribution to the functional differentiation of duplicated genes.

If the above-mentioned speculations are right, we have to ask very new and significant questions as follows. At the time of origin of life, there must have been the minimum set of ancestral minigenes that had been necessary for maintaining the life phenomena. What are they? How many minigenes had existed at that time? Of course, we cannot answer these questions immediately. However, there are some possibilities that we may be able to answer such difficult questions, because the DNA sequence data in various organisms are now accumulating with an enormous speed, and the nucleotide sequence data presently accumulated are freely available from the international DNA sequence data banks.

(*b*) *Making of multiple alignment database*

To approach the question about the minimum set of ancestral minigenes, the multiple alignment database was made from the presently available DNA sequence database (DDBJ release 12). This was done with the aim of classifying the available DNA sequences into appropriate groups on the sole basis of sequence homology. First, the nucleotide sequences for all possible coding regions were translated into the amino acid sequences. Then, a local homology search was conducted by use of 'FASTA' (Pearson & Lipman 1988) against all the amino acid sequence data by taking each entry sequence as a query sequence. Thus, the results of a local homology search were stored under the name of each entry sequence in the database called 'SODHO'.

For each entry in the 'SODHO' database, the global homology search was then conducted using Needleman and Wunsch's algorithm. Only the sequences whose homologies are quite high (namely, more than about 75%) are collected by this procedure. The collected homologous sequences are multiply aligned by conducting pairwise comparisons and adding the sequences to the alignment one by one. In this procedure, the results of multiple alignments depend on the order of adding sequences.

To avoid this difficulty, a new method for repeated constructions of phylogenetic trees was developed and it was used for making the multiple alignment. In the newly developed method, the tree was tentatively

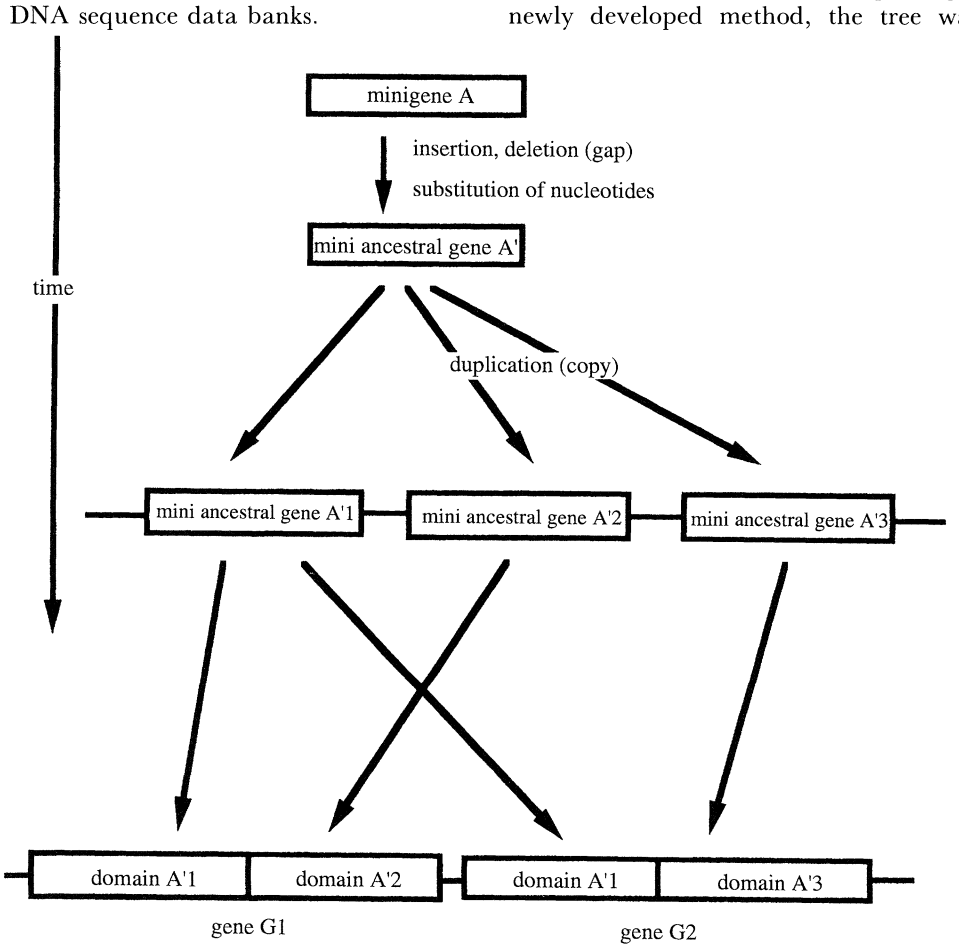

Figure 3. A possible feature of domain evolution. Functional domains are evolutionary units in which they change independently and dynamically in the evolutionary process.

constructed and sequence comparisons were performed following the order of the tree topology. Once a multiple alignment was tentatively made in this way, this procedure was repeated until the multiple alignment remains essentially unchanged. Thus, the final multiple alignment was adopted as one entry of the multiple alignment database called 'MultiAlign'. In this database, the multiple alignment data is stored under each of all entry sequences. The multiple alignment database, MultiAlign, has been published in the form of CD-ROM (Gojobori *et al.* 1993). The second version of MultiAlign has been released very recently.

(*c*) *Classification of homologous sequences*

MultiAlign contains many of the same gene set, because the multiple alignment was made under all entry sequences. Thus, the eventually same multiple alignments exist in MultiAlign, and we call them the 'overlapped alignment group'. For this reason, these alignment groups were excluded from the entire database in order to know the number of homologous gene sets in the DNA sequence database.

After conducting this procedure, 4 459 alignment groups were finally obtained as the non-overlapped alignment sets. Note that the alignment for each homologous gene group was made by use of the whole regions of amino acid sequences. In this procedure, functional domains have not been considered yet because these domains are, in general, parts of the entire amino acid sequences as discussed before. In the future, this number of the alignment sets should decrease drastically when the functional domains are taken into account.

(*d*) *Sequence motifs*

It is also important to extract 'sequence motifs' for each functional domain from the amino acid sequence data. When the amino acid sequences for homologous proteins are aligned with each other, it generally becomes clear that the sequences were strongly conserved at particular sites. Let us take the Kunitz-type protease inhibitor domain as an example. In the multiple alignment, all six cysteine residues are conserved at the respective sites where the distances between any pair of cysteines are almost constant. In addition, there are several other conserved amino acids such as a small stretch of tyrosine and glycines. This kind of characteristic configuration of amino acids is called 'sequence motif'. The sequence motif of the Kunitz-type protease inhibitor can be expressed as C[9X]C[17X]C[4X]YGGC[15X]C[3X]C, where C, Y, and G represent cysteine, tyrosine, and glycine, respectively, and any amino acid is denoted by X. For example, 5X represents a run of five amino acid sites where any amino acids can be taken.

Thus, extraction and classification of sequence motifs from the amino acid sequence data are essential for answering the questions of what the minimum set of ancestral minigenes have been and how many minigenes they were.

We thank Sir Walter Bodmer and Professor Peter Donnelly as well as the Royal Society for giving us a chance to write the present paper. We also thank Mrs Janet Clifford and Dr M. B. Goatly for their patience and encouragement. Without them, we could not have completed the present manuscript. The present study was supported in part by the Japanese Ministry of Education, Science and Culture.

REFERENCES

Doolittle, R.F. & Bork, P. 1993 Evolutionary mobile modules in proteins. *Scient. Am.* Oct, 50–56.

Gojobori, T., Moriyama, E.N., Ikeo, K. *et al.* 1993 Alignment of homologous amino acid sequences, SODHO Ver. 2.0. CD-ROM, Fujitsu Limited, Tokyo, Japan.

Ikeo, K., Takahashi, K. & Gojobori, T. 1991 Evolutionary origin of numerous Kringles in human and simian apolipoprotein(a). *FEBS Lett.* **287**, 146–148.

Ikeo, K., Takahashi, K. & Gojobori, T. 1992 Evolutionary origin of a Kunitz-type trypsin inhibitor domain inserted in the amyloid precursor protein of Alzheimer's disease. *J. molec. Evol.* **34**, 536–543.

Kang, J. & Mueller-Hill, B. 1989 The sequence of the two extra exons in rat preA4. *Nucl. Acids Res.* **17**, 2130–2130.

Kido, H., Fukutomi, A. & Katsunuma, N. 1991 Tryptase TL_2 in the membrane of human $T4^+$ lymphocytes is a novel binding protein of the V3 domain of HIV-1 envelope glycoprotein gp120. *FEBS Lett.* **287**, 233–236.

Masiakowski, P. & Caroll, R.D. 1992 A novel family of cell surface receptors with tyrosine kinase-like domain. *J. biol. Chem.* **267**, 26181–26190.

McLean, J.W., Tomlinson, J.E., Kuang, W.-J. *et al.* 1987 cDNA sequence of human apolipoprotein(a) is homologous to plasminogen. *Nature, Lond.* **330**, 132–137.

Miyazawa, K., Shinomura, T., Kitamura, A., Kondo, J., Morimoto, Y. & Kitamura, N. 1993 Molecular cloning and sequence analysis of the cDNA for a human serine protease responsible for activation of hepatocyte growth factor. *J. biol. Chem.* **268**, 10024–10028.

Muller-Hill, B. & Beyreuther, K. 1989 Molecular biology of Alzheimer's disease. *A. Rev. Biochem.* **58**, 287–307.

Nakamura, T., Nishizawa, T., Hagiya, M. *et al.* 1989 Molecular cloning and expression of human hepatocyte growth factor. *Nature, Lond.* **342**, 440–443.

Pathy, L. 1985 Evolution of the protease of blood coagulation and fibrinolysis by assembly from modules. *Cell* **41**, 657–663.

Pearson, W.R. & Lipman, D.J. 1985 Improved tools for biological sequence analysis. *Proc. natn. Acad. Sci. U.S.A.* **85**, 2444–2448.

Ponte, P., Gonzalez-DeWhitt, P., Schilling, J. *et al.* 1988 A new A4 amyloid mRNA contains a domain homologous to serine inhibitors. *Nature, Lond.* **333**, 525–527.

Saitou, N. & Nei, M. 1987 The neighbor-joining method: A new method for reconstructing phylogenetic trees. Estimation of average number of nucleotide substitutions when the rate of substitution varies with nucleotide. *Molec. Biol. Evol.* **4**, 406–425.

Wun, T.C., Kretzmer, K.K., Girard, T.J., Miletich, J.P. & Broze, G.J. 1988 Cloning and characterization of a cDNA coding for the lipoprotein associated coagulation inhibitor shows that it consists of three tandem Kunitz-type inhibitory domains. *J. biol. Chem.* **263**, 6001–6004.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES