PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Evolutionary Relationships among the Serpins

Craig J. Marshall

Phil. Trans. R. Soc. Lond. B 1993 **342**, 101-119
doi: 10.1098/rstb.1993.0141

Evolutionary relationships among the serpins

CRAIG J. MARSHALL

Department of Biochemistry, University of Otago, Box 56, Dunedin, New Zealand

SUMMARY

The serpins are a widely distributed group of serine proteinase inhibitors found in plants, birds, mammals and viruses. Despite the great evolutionary divergence of these organisms, their serpins are highly conserved, both in sequence and structurally. Amino acid sequences were aligned by a combination of automatic algorithms and by consideration of conserved structural elements in those serpins for which crystal structures exist. The program HOMED was used which allowed the alignment of amino acids to be simultaneously converted into the equivalently aligned nucleotide sequences. The aligned amino acids were used as the basis for superposition of the four known three-dimensional structures for which coordinates are available and compared with an optimal three-dimensional superposition in order to estimate the reliability of the sequence alignment. Phylogenetic relationships implied by these nucleotide sequence alignments were determined by the method of maximum parsimony. The proposed gene tree suggested that as much diversity existed between the plant serpin and mammalian serpins as was present among mammalian serpins and provided further evidence that the architecture of serpin molecules is highly constrained.

1. INTRODUCTION

Among the wide variety of proteinase inhibitors, the serpins (*ser*ine *p*roteinase *in*hibitor*s*) are characterized by a high molecular mass and a mechanism involving the separation of the amino acids about the scissile bond by about 70 Å† after proteolytic cleavage of the reactive centre (Bode *et al.* 1989). There is considerable evidence to suggest that a loop of peptide is exposed in the active serpin and is maintained in a configuration favourable for inhibition of a bound proteinase by the tendency of a strand to insert into a β-sheet (Carrell *et al.* 1991). Insertion of this strand proceeds to completion after cleavage of the active centre, which occurs upon the release of the proteinase, and results in a five-membered sheet converting to one of six strands. In this process a pair of parallel strands become an antiparallel trio upon insertion of the other strand (Loebermann *et al.* 1984; Huber & Carrell 1989; Engh *et al.* 1990).

The distribution of serpins ranges from protein Z in barley (Brandt *et al.* 1990), to chicken ovalbumin (McReynolds *et al.* 1978; Hunt & Dayhoff 1980; Stein *et al.* 1990), and to a variety of serpins found in mammals (Huber & Carrell 1989). In addition, there are a number of viral serpins that are probably derived from the genome of some host. The majority of serpins known are either viral or human; no serpins are known from fish, reptiles, or crustaceans; and only a few from insects (Kanost *et al.* 1989; Takagi *et al.* 1990), but it is probable that this represents the scope of investigations rather than the actual distribution of the family. Not all serpins act as proteinase inhibitors; ovalbumin apparently functioning as a storage protein in birds' eggs, angiotensinogen acting as a peptide hormone precursor, and barley protein Z demonstrating no inhibitory activity and possibly present in the barley endosperm as a storage protein.

The structures of four serpins are known: antitrypsin (Loebermann *et al.* 1984), ovalbumin (Wright *et al.* 1990; Stein *et al.* 1991), antichymotrypsin (Baumann *et al.* 1991), and plasminogen activator 1 (PAI1) (Mottonen *et al.* 1992). However, two of these structures, antitrypsin and antichymotrypsin, are of the cleaved molecule. The structures of both the cleaved and intact forms of ovalbumin are known and are essentially identical, but ovalbumin is not known to be inhibitory. Furthermore, the structure of PAI1 is of an intact but inactive form, as PAI1 loses activity relatively rapidly unless subjected to harsh conditions. Thus, although the structures of a number of members of the family are known, little direct information has been gathered about the nature of the inhibitory mechanism.

Examination of the serpin structures shows a high degree of architectural similarity (figure 1). The root mean square differences amongst the C_αs of serpins range from 0.67 Å between antitrypsin and antichymotrypsin, to 1.68 Å between antitrypsin and PAI1, and to 1.71 Å for the comparison of ovalbumin and PAI1. This conservation of structure implies that there are considerable constraints on the residues present at any particular part of the molecule, and that the serpins might be susceptible to phylogenetic analysis despite the large evolutionary distances involved. The structural similarity is reflected in the

† 1 Å = 10^{-10} m = 10^{-1} nm.

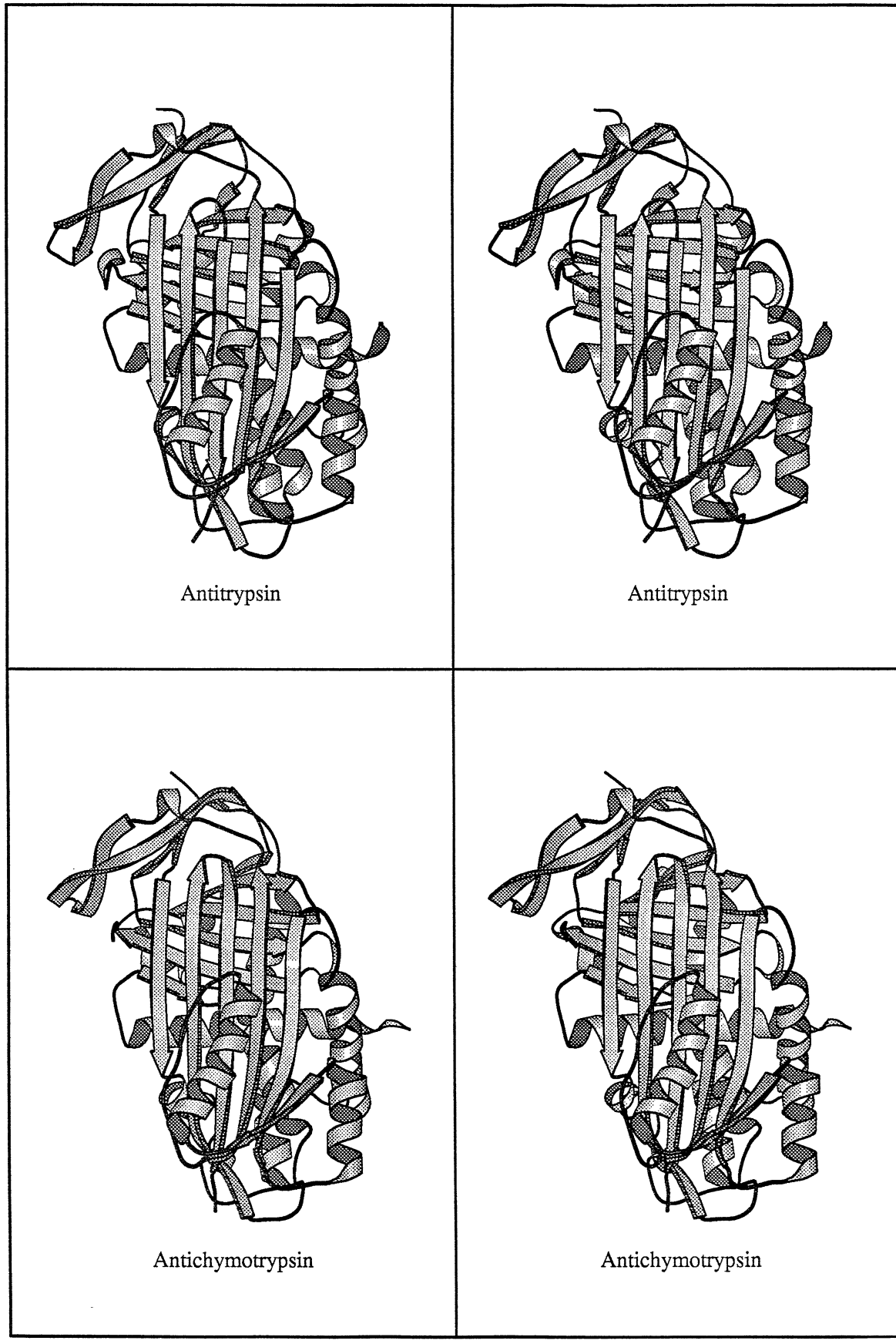

Figure 1. Cartoon representation of four serpins of known structure. Shown are the secondary structure elements of antitrypsin, antichymotrypsin, ovalbumin and plasminogen activator inhibitor I using the program Molscript (Kraulis 1991). The structures are each shown in an equivalent orientation. Some residues were disordered in the PAI1 coordinates and hence are absent in this figure.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

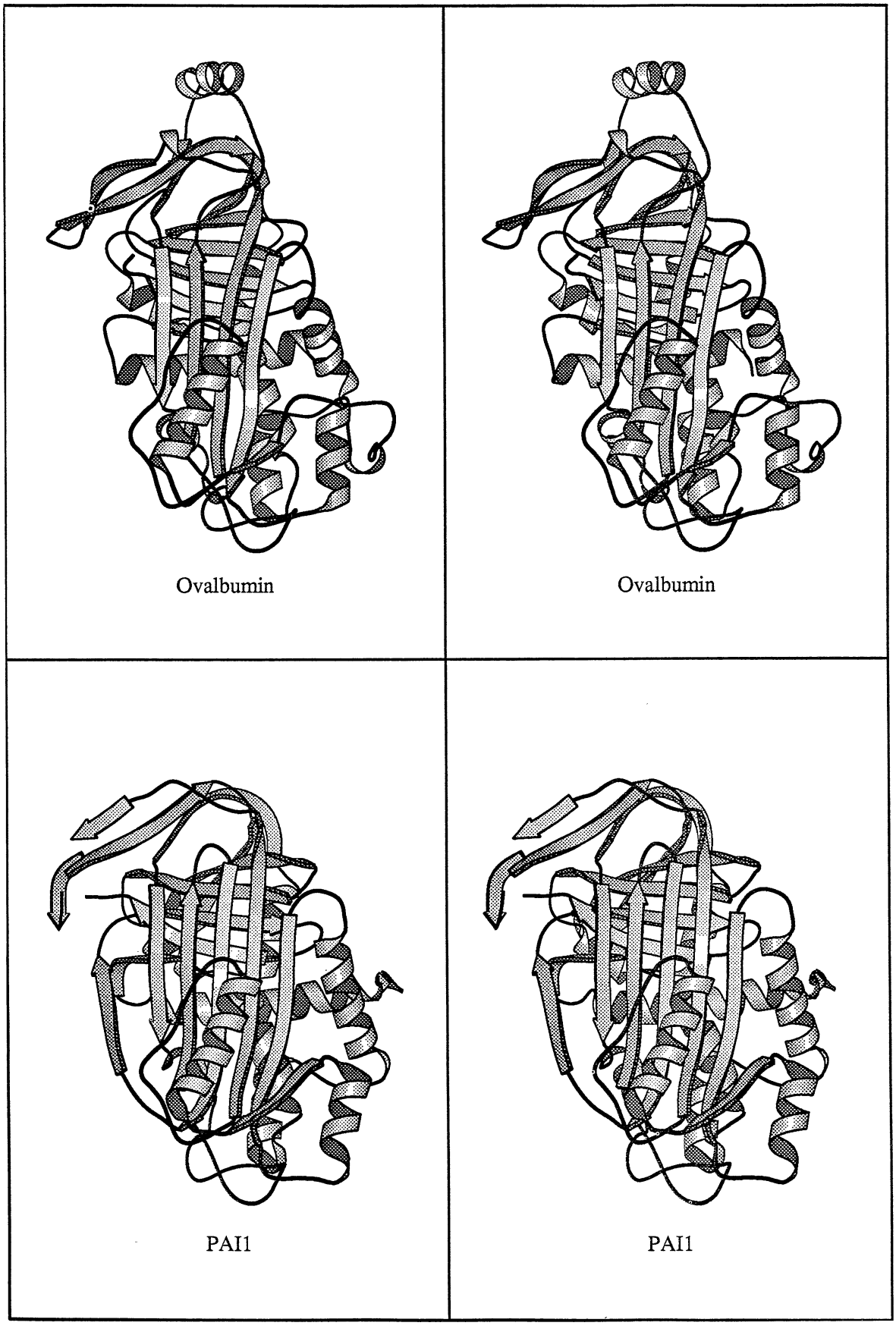

Figure 1. *Continued.*

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

sequences where strong conservation is found throughout the serpins. Simple assessments of sequence homology suggest conservative amino acid similarities of between 50% and 100% despite the considerable evolutionary distances between many of the distantly related sequences.

2. METHODS

(a) *Sequences*

All sequences used in this study were derived from either the EMBL 24 or GENBANK 66 nucleic acid databases except where specified otherwise. Where errors were detected the corrected version was used in this study and the error reported to the appropriate database operators. Only those serpins for which the complete nucleotide sequence was available were used. These are shown in table 1 along with the common names used in this paper, the reading frame used to produce the coding sequence, and the equivalent entries in the SWISSPROT protein database where appropriate.

(b) *Sequence alignment*

The translated nucleotide sequences were trimmed of the regions coding for peptide leader sequences, and in the cases of C_1-inhibitor (HSC1INHB) and antiplasmin (HSAPLA), regions corresponding to long amino or carboxy terminal extensions were removed. Initial alignment was done using the program AMPS (Barton & Sternberg 1987). An initial pairwise analysis using the Needleman and Wunsch algorithm (Needleman & Wunsch, 1970) was performed and used as the basis for subsequent multiple alignments. Most satisfactory results were obtained with a gap penalty of 12.0 and a constant of 8.0 (see Barton & Sternberg (1987) for details of these values).

The alignment produced was loaded into the program HOMED (Stockwell 1988). Regions of conserved secondary structure were determined from the superimposed structures of antitrypsin (Loebermann *et al.* 1984) and ovalbumin (Stein *et al.* 1990). The sequence alignment was refined manually; changes being made according to the principle that gaps and deletions in the sequence are more likely to be found in loops between regions of secondary structure than in α-helices or β structure (Lesk & Chothia 1980, 1982; Chothia & Lesk 1986; Lesk *et al.* 1986). However, relatively minor adjustments were required to produce a satisfactory alignment. Where appropriate, leader sequences and previously removed extensions were reintroduced to the alignment. No attempt was made to align the leader sequences as homologies in these regions were considered to be unrelated to their properties as serpins. This alignment is shown in figure 2. Also indicated in this alignment are the secondary structure assignments for antitrypsin and the antitrypsin numbering as well as those regions used in the phylogenetic analysis below.

The proportion of identical and conservative amino acid replacements was determined. The percentage of identical amino acids was calculated by assigning a value of 1 for a match or 0 for a nil match, and averaging the score over the length of the alignment. The distribution of identical amino acid replacements is shown in figure 3*a*. Conservative amino acid replacements were calculated by reference to a PAM 250 matrix adjusted by addition of a constant to remove all values less than or equal to zero. The value assigned to the pairing of residues i and j considered to be aligned was determined to be n_{ij}/n_{ii} where n_{ij} was the Dayhoff value for replacement of residue i by residue j and n_{ii} was the Dayhoff self-replacement value for residue i. The calculation was repeated for each aligned residue in the pair of sequences and the score averaged over the total length of the alignment. This procedure was repeated for each pair of aligned sequences (see Stockwell 1988) and the distribution of the values produced is shown in figure 3*b*.

To calculate theoretically expected values the relative frequencies of each amino acid in the aligned sequences was determined. For identical replacements the theoretical percentage identity can be determined by equation 1 where f_i is the relative frequency of the ith of the twenty amino acids.

$$I_{\text{predicted}} = \sum_{i=1}^{20} f_i^2. \qquad (1)$$

For conservative replacements the theoretical value is given by equation 2 where f_i and f_j are the relative frequency of each pair of amino acids, n_{ii}, n_{ij}, n_{ji} and n_{jj} are as above.

$$C_{\text{predicted}} = \sum_{i=1}^{20} \sum_{j=1}^{j \leqslant i} \frac{f_i \cdot f_j}{2} \left(\frac{n_{ij}}{n_{ii}} + \frac{n_{ji}}{n_{jj}} \right). \qquad (2)$$

Each pair of amino acids were considered as amino acid i replacing amino acid j and as j replacing i as a proportion of the self-replacement value (i.e. as i replacing i, and j replacing j), and the average of the two used as it was considered equally likely that i should replace j as j replace i. The theoretically expected scores are shown in figure 3*a*,*b*.

The amino acid alignment was converted to an alignment of nucleic acids using a new version of the program HOMED. In this version, nucleotide sequences comprise the database, but may be displayed and edited as either DNA or amino acid sequence.

(c) *Gene trees*

The aligned sequences were prepared for phylogenetic analysis by the removal of regions where many gaps and deletions were present, and by the removal of amino and carboxy terminal sequences. The major region removed from analysis was the area corresponding to the loop between the C and D helices of antitrypsin. Figure 2 indicates the regions of sequence used in phylogenetic analysis. The edited regions were written out from HOMED in a form suitable for analysis by the PHYLIP package of programs developed by Felsenstein (1988) and analysed as set out below.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Analysis by DNAPARS was done on 2000 bootstrap replacement sequence alignments produced by the program SEQBOOT, using the multiple sequence option of the program. During DNAPARS runs, a threshold level of 12 was chosen, whereby only the first twelve steps at a particular site in the analysis were considered in scoring, making the analysis intermediate between maximum parsimony and maximum likelihood (Felsenstein 1983, 1988). This has the advantage of reducing the tendency of maximum parsimony analysis to over-emphasize differences between sequences. The outgroup species was defined as barley protein Z (HVPROTZG) on the grounds that this was *a priori* the most evolutionarily distant sequence. Trees produced by these runs were weighted in inverse proportion to the number found in each individual analysis and the consensus sequence determined by the program CONSENSE from the package PHYLIP. The sequences were then fitted to this tree and the number of steps between each node determined and used for calculating the branch lengths. A similar analysis was done by adding the sequences to the tree in a random order which gave results very similar to that produced by the bootstrap method. An indication of the reliability of each of the branches of the tree was produced by calculating as a percentage the fraction of occurrences that a particular branch occurred in all the trees. These data are shown in figure 5 at each branch.

All analyses were performed on either a Digital DECStation 5000/200 running Ultrix V 4.2, or a MicroVAX II running the VMS operating system V 5.4.

(*d*) *Structural analysis*

The structures of antitrypsin (6API), ovalbumin (1OVA) from the Brookhaven Structural Database (Bernstein *et al.* 1977; Abola *et al.* 1987), PAI1 and antichymotrypsin were overlaid according to sequence similarity as determined above, or by comparison of similar secondary structural elements. The ovalbumin structure used was that of the intact form (Stein *et al.* 1990, 1991), as the cleaved and uncleaved forms are essentially identical.

The root mean square differences between C_α atoms of each structure after three-dimensional multiple least squares alignment were calculated by the program Quanta running on a Silicon Graphics Iris 4D25 computer. Superpositions were made by pairing residues considered to be structurally equivalent either by considering the sequence alignment in figure 2 or by examination of the known secondary structure of the molecules. The region corresponding to sheet s4A and the beginning of sheet s1C in antitrypsin (residues 345 to 365 inclusive) was not included in any of the alignments or calculations of fit as this region is known to vary widely in position in three of the four structures. The superpositions are shown in table 2 where the differences between the two different methods of superposition for each of the six possible pairings are shown.

3. RESULTS

(*a*) *Sequence alignment*

The aligned sequences are shown in figure 2. Sequences are labelled with the EMBL or GENBANK identifiers. A key to the common names can be found in table 1. Only those serpins for which the complete nucleotide sequence was available have been included in this study to avoid problems with aligning incomplete sequences and subsequent difficulties in determining phylogenetic relationships from incomplete data. Mouse contrapsin was one sequence excluded on these grounds. In addition, some sequences were not susceptible of analysis as although the protein sequence was available, the nucleotide sequence was not. Mouse angiotensinogen fell into this category. Indicated in figure 2 is the numbering and structural elements for antitrypsin (Loebermann *et al.* 1984), as are the regions selected for phylogenetic analysis.

Automatic multiple alignment of the serpins using AMPS (Barton & Sternberg 1987) was found to be significantly improved by the removal of amino and carboxy-terminal extensions. For example, the amino-terminal extension of C_1-inhibitor of about 200 amino acids containing a repeated glycosylation site, has a sequence quite unlike any of the other serpins. As this region is heavily glycosylated and appears to form a domain distinct from the serpin-like sequence, it is not surprising that its presence confounds attempts to align C_1-inhibitor with other serpins. Some serpins proved difficult to align and required considerable attention in some regions. In particular C_1-inhibitor caused difficulties which appear to be related to the presence of a repeated motif in the amino-terminal half of the sequence at about residue 100. Other proteins that were problematic included pig uteroferritin-associated protein and barley protein Z. However, even in these sequences, conserved motifs were present and allowed relatively unambiguous alignment of these regions at least.

The alignment produced by AMPS conformed well to other alignments (Huber & Carrell, 1989) and, with few exceptions, placed regions of sequence variation, especially of length, in the regions between structural elements. The notion of structurally conserved regions joined by loops of essentially variable structure and length is well established (Chothia & Lesk 1986; Lesk *et al.* 1986) and is a very useful consideration when aligning sequences where at least a few structures are known.

The sequence similarity expressed as percent identical amino acids and percent conservative amino acid replacements is shown in figure 3. The proportion of identical amino acids varies from about 12% to 100%, with the majority in the range 15% to 30% (figure 3*a*) suggesting rather low similarities in some instances. The calculated match for a set of sequences of this composition was 4.6%, less than half of the lowest measured value. Consideration of percent conservative amino acid changes indicates much stronger similarity with values typically in the range 55–75% (figure 3*b*), compared with a calculated value of 27%. Both distributions are skewed towards the upper

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Table 1. *The sequences used in the alignment*

(The common name for each sequence and the abbreviation used in the text, where applicable, is shown along with the EMBL or GENBANK identifier and the corresponding identifier from the SWISSPROT protein database where appropriate. The regions selected to produce the final reading frame are indicated.)

Common Name and abbreviation used in the text	EMBL 24 Entry	SWISSPROT Entry	Regions Selected for Reading Frame
Bovine plasminogen activator inhibitor 2 (bovine PAI2)	BTPAI1MR	PAI1$BOVINE	122..190, 191..1327
Chicken ovalbumin	CHKOVALM†	OVAL$CHICK	66..1226
Chicken gene Y	GGOVAY	OVAY$CHICK	1822..1989, 2535..2585, 2941..3069, 3917..4036, 4258..4398, 4480..4635, 5512..5910
Human antichymotrypsin	HUMA1ACM†	A1AC$HUMAN	15..86, 87..1310
Human antitrypsin	HSA1ATP	A1AT$HUMAN	7317..7387, 7388..7961, 9412..9682, 10939..11086, 11910..12098
Human angiotensinogen	HSANG	ANGT$HUMAN	40..138, 139..1494
Human antiplasmin	HSAPLA	A2AP$HUMAN	1..108, 109..1464
Human antithrombin	HSATIII	ANT3$HUMAN	47..1438
Human C1-inhibitor	HSC1INB	IC1$HUMAN	36..101, 102..1535
Human cortisol-binding globulin (CBG)	HSCBG	CBG$HUMAN	36..101, 102..1250
Human protein C inhibitor	HSCINHP	IPSP$HUMAN	47..103, 104..1264
Human glial-derived nexin (human GDN)	HSGDN	GDN$HUMAN	1..1191
Human heparin cofactor II (HCII)	HSHCII	HEP2$HUMAN	29..85, 86..1525
Human plasminogen activator inhibitor 2 (human PAI2)	HSPAI2	PAI2$HUMAN	56..1303
Human plasminogen activator inhibitor 1 (human PAI1)	HSPAIR	PAI1$HUMAN	127..195, 196..1332
Human thyroxine-binding globulin (TBG)	HSTBG	TBG$HUMAN	331..390, 391..1575
Barley protein Z	HVPROTZG‡	PRTZ$HORVU	Brandt *et al.* (1990)
Vaccinia serine proteinase inhibitor 1	M24217	SPI1$VACCV	927..1985
Vaccinia serine proteinase inhibitor 2	M24218	SPI2$VACCV	295..1320
Mouse antitrypsin	MMAAT	A1AT$MOUSE	9..1250
Manduca sexta proteinase inhibitor	MSPROI	SERA$MANSE	25..72, 73..1200
Mouse plasminogen activator inhibitor 2 (mouse PAI2)	MUSPAI2†	PAI2$MOUSE	13..1257
Sheep antitrypsin	OAPIA1AT	A1AT$SHEEP	6..77, 78..1253
Sheep uterine milk protein	OAUMPA	A33309	49..1338
Rabbit fibroma virus serpin	OCRFVHOM C*	YSER$RABBIT	365..1447 (of complement)
Pig antichymotrypsin	PIGA1ACM†		49..1326
Baboon antitrypsin	PPATRP	A1AT$PAPAN	3..1229
Cowpox (CPV-W2) gene	PXCPVWPV	HI38$COWPX	295..1320
Vaccinia B13R gene	PXVACB01		121..1158
Rat angiotensinogen	RNANG	ANGT$RAT	62..1494
Rat glial-derived nexin (rat GDN)	RNGDN	GDN$RAT	1..57, 58..1191
Rat serine proteinase inhibitor 1	RNSEPI1		130..1338
Rat serine proteinase inhibitor 2	RNSEPI2		123..1343
Rat serine proteinase inhibitor 3	RNSEPI3		129..1355
Rat spi-1 serpin	RNSPI1	SI1$RAT	77..1045
Rat spi-2 serpin	RNSPI2	SI20$RAT	60..1320
Rat spi-2.3 serpin	RNSPI23	SI23$RAT	82..1329
Pig uteroferrin-associated protein	SSUFBP	UFBP$PIG	64..1317

†Entry taken from GENBANK 66. ‡Entry modified from HVPROTZ according to Brandt *et al.* (1990). *The "C" indicates that the complement of the sequence entry OCRFVHOM was used in this study.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES


```
                  1        11        21        31        41        51        61        71        81        91       101       111
AT Numbering                                                                                                                  1
AT Structure
RNSEPI1                                                                                                                MAGICPAVLCDGT
RNSEPI2                                                                                                                MAGICPAVLCDGI
RNSEPI3                                                                                                                MDGIGSALLSFPDCI
RNSPI2                                                                                                              MAFIAALGLLMAGICPAVLCDGI
RNSPI23                                                                                                             MAFIAALGLLMAGICPAVLCDGI
HSA1ACM                                                                                                           MERMLPLLALGLLAAGFCPAVLCHPNS
PIGA1ACM                                                                                                          MRAEGMSLFLALGLLVAGLCSRVHCVPADD
HSA1ATP                                                                                                           MPSSVSWGILLLAGLCCLVPVSLAEDPQ
PPATRP                                                                                                              MLLLAGLCCLLPGSLAEDPQ
MMAAT                                                                                                               MTPSISWGLLLLAGLCCLVPSF
OAPIA1T                                                                                                           MALSITRGLLLLAALCCLAPTSLAGV
HSCBG                                                                                                               MPLLLYTCLLWLPTSGL
HSCINHP                                                                                                             MQLFLLLCLVLLSPQGASL
HSTBG                                                                                                               MSPFLYLVLLVLGLHATIHC
HSHCII                                MKHSLNALLIFLIITSAWGGSKGPLDQLEKGGETAQSADPQWEQLNNKNLSMPLLPADFHKENTVTNDWIPEGEEDDDYLDLEKIFSEDDDYIDIVDSLSVSPT
HSANG                                                 MRKRAPQSEMAPAGVSLRATILCLLAWAGLAAGDRVYIHPFHLVIHNESTCEQLAKANAGKPKDPTFIPAPIQAKTSPVDEKAL
RNANG                                                                 MTPTGAGLKATIFCILTWVSLTAGDRVYIHPFHLLYYSKSTCAQLENPSVETLPEPTFEPVPIQAKTSPVDEKTL
HSATIII                                                                                 MYSNVIGTVTSGKRKVYLLSLLLIGFWDCVTCHGSPVDICTAKPRDIPMNPMCIYRSPEKK
HSPAI2
MOUSE PAI2                                                                                        ,
HVPROTZG
CHKOVALM
GGOVAY
HSAPLA                                                                                                  MLWGLLVLSWSCLQGPCSVFSPVSAMEPLGRQLTSGPNQEQVSPLTLLKLGNQEPGGQ
HSC1INHB      MASRLTLLTLLLLLLAGDRASSNPNATSSSSQDPESLQDRGEGKVATTVISKMLFVEPILEVSSLPTTNSTTNSATKITANTTDEPTTQPTTEPTTQPTIQPTQPTTQLPTDSPTQPTTG
MSPROI                                                                                                                        MKI
M24217
M24218
PXCPVWPV
PXVACB01
OCRFVHOM C
HSGDN                                                                                                                      MNWHL
RRGDN                                                                                                                      MNWHF
BTPAI1MR                                                                                                                 MRMSPVFAC
HSPAIR                                                                                                                   MQMSPALTC
OAUMPA                                                                                                            MSHRRMQLALSLVFILCGLFNSIFCEKQ
SSUFBP                                                                                                            MSHGKMPLVLSLVLILCGLFNSISCEKQ
Phylip
AT Structure
AT Numbering                                                                                                                  1
```

Figure 2. Alignment of serpins. Serpin names are as given in table 1. The numbering and secondary structure assignments for antitrypsin are derived from Huber & Carrell (1989) and are shown in the lines labelled AT Numbering and AT Structure. α Helices are designated by the character ^and strands comprising β-sheet by −. The symbol | | designates boundaries between two similar structures. An arbitrary gap marked by the symbols > and < was introduced between the P_1 and P_1' residues of antitrypsin. The regions of the alignment used for phylogenetic analysis are denoted by an + in the line labelled Phylip.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Figure 2. *Continued.*

```
                  121       131       141       151       161       171       181       191       201       211       221       231
AT Numbering          10        20        30        40        50        60        70        80
AT Structure                 ^^^^^^^^^^^^^hA^^^^^ ^^^^^ ^   -s6b-^^^^^^^hB^^^^^^^||^^^^^^hC^^^^^ |                                 hC1||^
RNSEPI1       LGRDTLSHEDHGKGRQLHSLTLASSNTDFALSLY KKLAL RNPDKNVVFSPLSISAALTILSLGAKDSTMEEILEGLKFN                                 LTEITE
RNSEPI2       LGRDTLPHEDQGKGRQLHSLTLASINTDFTLSLY KKLAL RNPDKNVVFSPLSISAALAILSLGAKDSTMEEILEVLKFN                                 LTEITE
RNSEPI3       LGEDTLFHEDQDKGTQLDSLTLASINTDFAFSLY KKLAL RNPHKNVVFSPLSISAALAVVSLGAKGSSMEEILEGLKFN                                 LTETPE
RNSPI2        LGRDTLPHEDQGKGRQLHSLTLASINTDFTLSLY KKLAL RNPDKNVVFSPLSISAALAILSLGAKDSTMEEILEGLKFN                                 LTEITE
RNSPI23       LGRDTLPHEDQGKGRQLHSLTLASINTDFTLSLY KKLAL RNPDKNVVFSPLSISAALAILSLGAKDSTMEEILEGLKFN                                 LTEITE
HSA1ACM       PLDEENLTQENQDRGTHVDLGLASANVDFAFSLY KQLVL KALDKNVIFSPLSISTALAFLSLGAHNTTLTEILKGLKFN                                 LTETSE
PIGA1ACM      LASKIVTLKDQIKKLPVHNTAVVSSNTDFAFSLY KQLSL TNRHKNIIFSPLSVSMALFLSLGARGPTLTELLKASSSCH                                 RDSLR
HSA1ATP       GDAAQKTDTSHHDQDHPTFNKITPNLAEFAFSLY RQLAH QSNSTNIFFSPVSIATAFAMLSLGTKADTHDEILEGLNFN                                 LTEIPE
PPATRP        GDAAQKTDTPPHDQNHPTLNKITPSLAEFAFSLY RQLAH QSNSTNIFFSPVSIATAFAMLSLGTKADTHSEILEGLNFN                                 LTEIPE
MMAAT         LAEDVQETDTSQKDQSPASHEIATNLGDFAISLY RELVH QSNTSNIFFSPVSIATAFAMLSLGSKGDTHTQILEGLQFN                                 LTQTSE
OAPIA1T       LQGHAVQETDDTAHQEAACHKIAPNLANFAFSIY HKLAH QSNTSNIFFSPVSIASAFAMLSLGAKGNTHTEILEGLGFN                                 LTELAE
HSCBG         WTVQAMDPNAAYVNMSNHHRGLASANVDFAFSLY KHLVA LSPKKNIFISPVSISMALAMLSLGTCGHTRAQLLQGLGFN                                 LTERSE
HSCINHP       HRHHPREMKKRVEDLHVGATVAPSSRRDFTFDLY RALAS AAPSQNIFFSPVSISMSLAMLSLGAGSSTKMQILEGLGLN                                 LQKSSE
HSTBG         ASPEGKVTACHSSQPNATLYKMSSINADFAFNLY RRFTV ETPDKNIFFSPVSISAALVMLSFGACCSTQTEIVETLGFN                                 LTDTPM
HSHCII        DSDVSAGNILQLFHGKSRIQRLNILNAKFAFNLY RVLKDQVNTFDNIFIAPVGISTAMGMISLGLKGETHEQVHSILHFK                                 DFVNASSKYEI
HSANG         QDQLVLVAAKLDTEDKLRAAMVGMLANFLGFRIYGMHSELWGVVHGATVLSPTAVFGTLASLYLGALDHTADRLQAILGVP                                 WKDKNCTSRLDAHKVL
RNANG         RDKLVLATEKLEAEDRQRAAQVAMIANFMGFRMY KMLSEARGVASGAVLSPPALFGTLVSFYLGSLDPTASQLQVLLGVP                                 VKEGDCTSRLDGHKVL
HSATIII       ATEDEGSEQKIPEATNRRVWELSKANSRFATTFY QHLADSKNDNDNIFLSPLSISTAFAMTKLGACNDTLQQLMEVFKFDT                                 ISEKTS
HSPAI2                          MEDLCVANTLFALNLF KHLAK ASPTQNLFLSPWSISSTMAMVYMGSRGSTEDQMAKVLQFNEVGANAVTPMTPENFTSCGFMQQIQKGSYPDAILQAQAA
MOUSE PAI2                      MEELSMANTMFALNLL KQIEK SNSTQNIFISPWSISSTLAIVLLGAGGNTEQQMAKVLQFNEIGSYGITTRNPENFSGCDFAQQIQKENYPSAILQAQAG
HVPROTZG              MATTLATDVRLSIAHQTRFALRLRSA ISSNP ERAAGNVAFSPLSLHVALSLITAGAAATRDQLVAILGDGG                                 AGDAKEL
CHKOVALM                        MGSIGAASMEFCFDVF KELKV HHANENIFYCPIAIMSALAMVYLGAKDSTRTQINKVVRFD                                 KLPGFGDSIEAQCGTS
GGOVAY                          MDSISVTNAKFCFDVF NEMKV HHVNENILYCPLSILTALAMVYLGARGNTESQMKKVLHFD                                 SITGAGSTTDSQCGSS
HSAPLA        TALKSPPGVCSRDPTPEQTHRLARAMMAFTADLF SLVAQ TSTCPNLILSPLSVALALSHLALGAQNHTLQRLQQVLHAG
HSC1INHB      SFCPGPVTLCSDLESHSTEAVLGDALVDFSLKLY HAFSAMKKVETNMAFSPFSIASLLTQVLLGAGQNTKTNLESILSYP
MSPROI        IMCIFGLAALAMAGETDLQKILRESNDQFTAQMF SEVVK ANPGQNVVLSAFSVLPPLGQLALASVGESHDELLRALALP                                 N
M24217                                  MDIF KELIL KHTDENVLISPVSILSTLSILNHGAAGSTAEQLSKYIE
M24218                                  MDIF REIAS SMKGENVFISPASISSVLTILYYGANGSTAEQLSKYVE
PXCPVWPV                                MDIF REIAS SMKGENVFISPPSISSVLTILYYGANGSTAEQLSKYVE
PXVACB01                                MDIF REIAS SMKGENVFISPASISSVLTILYYGANGSTAEQLSKYVE
OCRFVHOM C                        MFNVVRVRDIGLWTF RYVY   NESDNVVFSPYGLTSALSVLRIAAGGNTKREIDVPESV
HSGDN         PLFLLASVTLPSICSHFNPLSLEELGSNTGIQVF NQIVK SRPHDNIVISPHGIASVLGMLQLGADGRTKKQLAMVMRYG                                 V
RRGDN         PFFILTTVTLSSVYSQLNSLSLEELGSDTGIQVF NQIIK SQPHENVVISPHGIASILGMLQLGADGRTKKQLSTVMRYN                                 V
BTPAI1MR      LALGLALIFGEGSASYQPQSAAASLATDFGVKVF QQVVR ASKDRNVVFSPYGVASVLAMLQLTTGGETRQQIQEAMQFK                                 IEE
HSPAIR        LVLGLALVFGEGSAVHHPPSYVAHLASDFGVRVF QQVAQ ASKDRNVVFSPYGVASVLAMLQLTTGGETQQQIQAAMGFK                                 IDD
OAUMPA        QHSQQHANLVLLKKISAFSQKMEAHPKAFAQELF KALIA ENPKKNIIFSPAAMTITLATLSLGIKSTMSTNHPEDLELELK                                 L
SSUFBP        QTSPKTITPVSFKRIAALSQKMEANYKAFAQELF KTLLI EDPRKNMIFSPVSISISLATLSLGLRSATRTNAIDVLDVALK                                 NLAV
Phylip                          ++++++++++++++++ +++++ ++++++++++++++++++++++++++++++++++++
AT Structure                 ^^^^^^^^^^^^^hA^^^^^ ^^^^^ ^   -s6b-^^^^^^^hB^^^^^^^||^^^^^^hC^^^^^ ^                                 hC1||^
AT Numbering          10        20        30        40        50        60        70        80
```

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Figure 2. *Continued.*

```
                241       251       261       271       281       291       301       311       321       331       341       351
AT Numbering 90        100       110       120       130       140       150       160       170       180       190       200
AT Structure  ^^^ ^^hD^^^^^^^^^    -----s2a-----     ^^^^^^hE^^^^||- s1A--   ^^^^^^^hF^^^^^^^^^                ------s3A-----     ^^
RNSEPI1       EEI HQGFGHLLQRLSQ PEDQVEINTGSALFIDKEQPILSEFQEKTRALYQAE AFIADF KQPNEAKKLINDYVSNQTQGKIAELF   SDLEERTSMVLVNYLLFKGKWKVPFNPND
RNSEPI2       EEIHHQGFGHLLQRLSQ PEDQAEINTGSALFIDKEQPILSEFQEKTRALYQAE AFVADF KQCNEAKKFINDYVSNQTQGKIAELF   SELDERTSMVLVNYLLFKGKWKVPFNPND
RNSEPI3       TEI HRGFGHLLQRLSQ PRDEIQISTGNALFIEKRLQVLAEFQEKAKALYQAE AFTADF QQSREAKKLINDYVSKQTQGKIQGLI   TNLAKKTSMVLVNYIYFKGKWKVPFDPRD
RNSPI2        EEI HQGFGHLLQRLSQ PEDQAEINTGSALFIDKEQPILSEFQEKTRALYQAE AFVADF KQCNEAKKFINDYVSNQTQGKIAELF   SELDERTSMVLVNYLLFKGKWKVPFNPND
RNSPI23       EEI HQGFGHLLQRLSQ PEDQAEINTGSALFIDKEQPILSEFQEKTRALYQAE AFVADF KQCNEAKKFINDYVSNQTQGKIAELF   SELDERTSMVLVNYLLFKGKWKVPFNPND
HSA1ACM       AEI HQSFQHLLRTLNQ SSDELQLSMGNAMFVKEQLSLLDRFTEDAKRLYGSE AFATDF QDSAAAKKLINDYVKNGTRGKITDLI   KDPDSQTMMVLVNYIFFKAKWEMPFDPQD
PIGA1ACM      QKS TQGFQHLRGRSIR PSNLLQLRLGNAMFIDEQLELVDKFVQDARELYHSE AFPTNF QDLEAARRLINDYVKNKTEGKIVELI   KRLIPVTEAVLVNYIYFKAQWMTHFDPNL
HSA1ATP       AQI HEGFQELLRTLNQ PDSQLQLTTGNGLFLSEGLKLVDKFLEDVKKLYHSE AFTVNF GDTEEAKKQINDYVEKGTQGKIVDLV   KELDRDTVFALVNYIFFKGKWERPFEVKD
PPATRP        AQV HEGFQELLRTLNK PDSQLQLTTGNGLFLNKSLKVVDKFLEDVKNLYHSE AFSVNF EDTEEAKKQINNYVEKGTQGKVVDLV   KELDRDTVFALVNYIFFKGKWERPFEVEA
MMAAT         ADI HKSFQHLLQTLNR PDSELQLSTGNGLFVNNDLKLVEKFLEEAKNHYQAE VFSVNF AESEEAKKVINDFVEKGTQGKIVEAV   KELDQDTVFALGNYILFKGKWKKPFDPEN
OAPIA1T       AEI HKGFQHLLHTLNQ PNHQLQLTTGNGLFINESAKLVDTFLEDVKNLHHSK AFSINF RDAEEAKKKINDYVEKGSHGKIVDLV   KDLDQDTVFALVNYISFKGKWEKPFEVEH
HSCBG         TEI HQGFQHLHQLFAK SDTSLEMTMGNALFLDGSLELLESFSADIKHYYESE VLAMNF QDWATASRQINSYVKNKTQGKIVDLF   SGLDSPAILVLVNYIFFKGTWTQPFDLAS
HSCINHP       KEL HRGFQQLLQELNQ PRDGFQLSLGNALFTDLVVDLQDTFVSAMKTLYLAD TFPTNF RDSAGAMKQINDYVAKQTKGKIVDLL   KNLDSNAVVIMVNYIFFKAKWETSFNHKG
HSTBG         VEI QHGFQHLICSLNF PKKELELQIGNALFIGKHLKPLAKFLNDVKTLYETE VFSTDF SNISAAKQEINSHVEMQTKGKVVGLI   QDLKPNTTMVLVNYIHFKAQWANPFDPSK
HSHCII        TTI HNLFRKLTHRLFR RNFGYTLRSVNDLYIQKQFPILLDFRTKVREYYFAE AQIADF  SDPAFISKTNNHIMKLTKGLIKDAL   ENIDPATQMMILNCIYFKGSWVNKFPVEM
HSANG         SAL QAVQGLLVAQGRADSQAQLLLSTVVGVFTAPGLHLKQPFVQGLALYTPVVLPRSLDF TELDVAAEKIDRFMQAVTGWKTGCSL   MGASVDSTLAFNTYVHFQGKMK GFSLLA
RNANG         TAL QAVQGLLVTQGGSSSQTPLLQSTVVGLFTAPGLRLKQPFVESLGPFTPAIFPRSLDLSTDPVLAAQKINRFVQAVTGWKMNLPL   EGVSTDSTLFFNTYVHFQGKMR GFSQLT
HSATIII       DQI HFFFAKLNCRLYRKANKSSKLVSANRLFGDKSLTFNETYQDISELVYGAK LQPLDFKENAEQSRAAINKWVSNKTEGRITDVI PSEAINELTVLVLVNTIYFKGLWKSKFSPEN
HSPAI2        DKI HSSFRSLSSAINA STGDYLLESVNKLFGEKSASFREEYIRLCQKYYSSE PQAVDFLECAEEARKKINSWVKTQTKGKIPNLL PEGSVDGDTRMVLVNAVYFKGKWKTPFEKKL
MOUSE PAI2    DKI HSAFSSLSSTINT PQGDYLLESANKLFGEKSARFKEEYIQLSKKYYSTE PEAVDFLECAEEAREKINSWVKTQTKGEIPNLL PEGSVDEDTKMVLVNAVYFKGKWKTPFEKKL
HVPROTZG      NAL AEQVVQFVLANES STGGPRIAFANGIFVDASLSLKPSFEELAVCQYKAK TQSLDFQHKTLEAVGQVNSWVEQVTTGLIKQIL PPGSVDNTTKLILGNALYFKGAWDQKFDESN
CHKOVALM      VNV HSSLRDILNQITK PNDVYSFSLASRLYAEERYPILPEYLQCVKELYRGG LEPINFQTAADQARELINSWVESQTNGIIRNVL QPSSVDSQTAMVLVNAIVFKGLWEKTFKDED
GGOVAY        EYV HNLFKELLSEITR PNATYSLEIADKLYVDKTFSVLPEYLSCARKFYTGG VEEVNFKTAAEEARQLINSWVEKETNGQIKDLL VSSSIDFGTTMVFINTIYFKGIWKIAFNTED
HSAPLA          S GPCLPHLLSRLCQ DLGPGAFRLAARMYLQKGFPIKEDFLEQSEQLFGAK  PVSLT GKQEDDLANINQWVKEATEGKIQEFL   SGLPEDTVLLLLNAIHFQGFWRNKFDPSL
HSC1INHB        KDFTCVHQALK GFTTKGVTSVSQIFHSPDLAIRDTFVNASRTLYSSS  PRVLS NNSDANLELINTWVAKNTNNKISRLL   DSLPSDTRLVLLNAIYLSAKWKTTFDPKK
MSPROI        DNV TKDVFADLNRGVR AVKGVDLKMASKIYVAKGLELNDDFAAVSRDVFGSE VQNVDF VKSVEAAGAINKWVEDQTNNRIKNLV DPDALDETTRSVLVNAIYFKGSWKDKFVKER
M24217        NMN ENTPDDNNDMDVD IPYCATLATANKIYGSDSIEFHASFLQKIKDDFQT     VNF NNANQTKELINEWVKTMTNGKINSLL  TSPLSINTRMTVVSAVHFKAMWKYPFSKHL
M24218              KEENMDKV    SVSFKSINKVYGRYSAVFKDSFLRKIGDKFQT     VDF  TDCRTIDAINKCVDIFTEGKINPLL  DEPLSPDTCLLSISADVFKAKWLTPFEKEF
PXCPVWPV            KEADKNKD    DISFKSMNKVYGRYSAVFKDSFLRKIGDNFQT     VDF  TDCRTVDAINKCVDIFTEGKINPLL  DEPLSPDTCLLAISAVYFKAKWLMPFEKEF
PXVACB01            KEENMDKV SAQNISFKSINKVYGRYSAVFKDSFLRKIGDKFQT     VDF  TDCRTIDAINKCVDIFTEGKINPLL  DEPLSPDTCLLAISAVYFKAKWLTPFEKEF
OCRFVHOM C                  VEDSDAFLALRELFVDASVPLRPEFTAEFSSRFNTS VQRVTF  NSENVKDVINSYVKDKTGGDVPRVL  DASLDRDTKMLLLSSVRMKTSWRHVFDPSF
HSGDN         NGV GKILKKINKAIVS KKNKDIVTVANAVFVKNASEIEVPFVTRNKDVFQCE VRNVNF EDPASACDSINAWVKNETRDMIDNLLSPDLIDGVLTRLVLVNAVYFKGLWKSRFQPEN
RRGDN         NGV GKVLKKINKAIVS KKNKDIVTVANAVFVRNGFKVEVPFAARNKEVFQCE VQSVNF QDPASACDAINFWVKNETRGMIDNLLSPNLIDSALTKLVLVNAVYFKGLWKSRFQPEN
BTPAI1MR      KGM APAFHRLYKELMG PWNKDEISTADAIFVQRDLELVHGFMPNFFRLFRTT VKQVDF SEVERARFIVNDWVKRHTKGMISDLL GEGAVDQLTRLVLVNALYFNGQWKMPFPESN
HSPAIR        KGM APALRHLYKELMG PWNKDEISTTDAIFVQRDLKLVQGFMPHFFRLFRST VKQVDF SEVERARFIINDWVKTHTKGMISNLL GKGAVDQLTRLVLVNALYFNGQWKTPFPDSS
OAUMPA        LDA HKCLHHLVHLGRE LVKQKQLRHQDILFLNSKMMANQMLLHQIRKLQKMD IQMIDF SDTEKAKKAISHHVAEKTHTKIRDLI   TDLNPETILCLVNHIFFKGILKRAFQPNL
SSUFBP        MLM AQAPTALLEIVHE LVNR TAKHQDILI  DRTEMNQMFLKEIDRYIKMD IQMIDF KDKEKTKKAINQFVADKIDKKAKNLI   THLDPQTLLCLVNYIFFKGILERAFQTNL
Phylip            ++++++++++++ ++++++++++++++++++++++++++++++++++++++ ++++++  ++++++++++++++++++++++++   ++++++++++++++++++++++++++++
AT Structure  ^^^ ^^hD^^^^^^^^^    -----s2a-----     ^^^^^^hE^^^^||- s1A--   ^^^^^^^hF^^^^^^^^^                ------s3A-----     ^^
AT Numbering 90        100       110       120       130       140       150       160       170       180       190       200
```

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Figure 2. *Continued.*

```
          361       371       381       391       401       411       421       431       441       451       461       471
AT Numbering        210       220       230       240       250       260       270       280       290       300
AT Structure  hF1-- s4C--||---s3C----------  ^hF2^-----s2B--  ---s3B---    ^^hG^^  ^^^^^hH^^  ^^    --s2C-||-s6A  -----^^^hI^^^
RNSEPI1       TFESE FYLDEKRSVKVPMMKIKEVTTPYVR DEELSCSVLELKYTG NASALFILPDQ    GKMQQVESSLQPETLKKWK  DSLIPRIINDLRMPKFSIS  TDYSLKEVLPELG
RNSEPI2       TFESE FYLDEKRSVKVPMMKIKDLTTPYIR DEELSCSVLELKYTG NASALFILPDQ    GKMQQVESSLQPETLKKWK  DSLRPRIISELRMPKFSIS  TDYNLEEVLPELG
RNSEPI3       TFQSE FYSGKRRSVKVPMMKLEDLTTPYVR DEELNCTVVELKYTG NASALFILPDQ    GKMQQVEASLQPETLRRWK  DSLRPSMIDELYLPKFSIS  ADYNLEDVLPELG
RNSPI2        TFESE FYLDEKRSVKVPMMKIKDLTTPYIR DEELSCSVLELKYTG NASALFILPDQ    GKMQQVESSLQPETLKKWK  DSLRPRIISELRMPKFSIS  TDYNLEEVLPELG
RNSPI23       TFESE FYLDEKRSVKVPMMKIKDLTTPYIR DEELSCSVLELKYTG NASALFILPDQ    GKMQQVESSLQPETLKKWK  DSLRPRIISELRMPKFSIS  TDYNLEEVLPELG
HSA1ACM       THQSR FYLSKKKWVMVPMMSLHHLTIPYFR DEELSCTVVELKYTG NASALFILPDQ    DKMEEVEAMLLPETLKRWR  DSLEFREIGELYLPKFSIS  RDYNLNDILLQLG
PIGA1ACM      TTEAD FYVSKNRTVRVPMMAIKNLTVPYFR DEELACTVVELPYTS NDSTLFILPDD    GRMAAVEAKLLPETLRRWR  DSLQPRRIEALHLPRFSIS  SDYKLQEILPRLG
HSA1ATP       TEEED FHVDQVTTVKVPMMKRLGMFNIQH  CKKLSSWVLLMKYLG NATAIFFLPDE    GKLQHLVNELTHDIITKFL  ENEDRRSASLHLPKLSIT  GTYDLKSVLGQLG
PPATRP        TEEED FHVDQATTVKVPMMRRLGMFNIYH  CEKLSSWVLLMKYLG NATAIFFLPDE    GKLQHLENELTHDIITKFL  ENENRRSANLHLPKLAIT  GTYDLKTVLGHLG
MMAAT         TEEAE FHVDKSTTVKVPMMTLSGMLDVHH  CSTLSSWVLLMDYAG NASAVFLLPED    GKMQHLEQTLNKELISKIL  LNRRRRLVQIHIPRLSIS  GEYNLKTLMSPLG
OAPIA1T       TTERD FHVNEQTTVKVPMMNRLGMFDLHY  CDKLASWVLLLDYVG NVTACFILPDL    GKLQQLEDKLNNELLAKFL  EKKYASSANLHLPKLSIS  ETYDLKTVLGELG
HSCBG         TREEN FYVDETTVVKVPMMLQSSTISYLH  DSELPCQLVQMNYVG NGTVFFILPDK    GKMNTVIAALSRDTINRWS  AGLTSSQVDLYIPKVTIS  GVYDLGDVLEEMG
HSCINHP       TQEQD FYVTSETVVRVPMMSREDQYHYLL  DRNLSCRVVGVPYQG NATALFILPSE    GKMQQVENGLSEKTLRKWL  KMFKKRQLELYLPKFSIE  GSYQLEKVLPSLG
HSTBG         TEDSSSFLIDKTTTVQVPMMHQMEQYYHLV  DMELNCTVLQMDYSK NALALFVLPKE    GQMESVEAAMSSKTLKKWN  RLLQKGWVDLFVPKFSIS  ATYDLGATLLKMG
HSHCII        THNHN FRLNEREVVKVSMMQTKGNFLAAN  DQELDCDILQLEYVG GISMLIVVPHK    MSGMKTLEAQLTPRVVERWQ  KSMTNRTREVLLPKFKLE  KNYNLVESLKLMG
HSANG         EP QE FWVDNSTSVSVPMLSGMGTFQHWS  DIQDNFSVTQVPFTE SACLLLIQPHY    ASDLDKVEGLTFQQNSLNWM  KKLSPRTIHLTMPQLVLQ  GSYDLQDLLAQAE
RNANG         GL HE FWVDNSTSVSVPMLSGTGNFQHWS  DAQNNFSVTRVPLGE SVTLLLIQPQ     CASDLDRVEVLVFQHDFLTWI  KNPPPRAIRLTLPQLEIR  GSYNLQDLLAQAK
HSATIII       TRKEL FYKADGESCSASMMYQEGKFRYRR  VAEGTQVLELPFKGD DITMVLILPKP    EKSLAKVEKELTPEVLQEWL  DELEEMMLVVHMPRFRIE  DGFSLKEQLQDMG
HSPAI2        NGLYP FRVNSAQRTPVQMMYLREKLNIGY  IEDLKAQILELPYAG DVSMFLLLPDEIADVSTGLELLESEITYDKLNKWTS KDKMAEDEVEVYIPQFKLE  EHYELRSILRSMG
MOUSE PAI2    NGLYP FRVNSHESIPVQMMFLHAKLNIGY  IKDLKTQILELPHTG NISMLLLLPDEIEDASTGLELLESEINFANFNKWIS KDTLDEDDVVVYIPKFKLA  QSYELKSILQSMG
HVPROTZG      HKCDS FHLLDGSSIQTQFMSSTKKQYISSS    DNLKVLKLPYAKGHDKRQFSMYILLPGAQDGLWSLAKRLSTEPEFIENH IPKQTVEVGRFQLPKFKIS  YQFEASSLLRALG
CHKOVALM      TQAMP FRVTEQESKPVQMMYQIGLFRVASM   ASEKMKILELPFASGTMSMLVLLPDE    VSGLEQLESIINFEKLTEWTS SNVMEERKIKVYLPRMKME  EKYNLTSVLMAMG
GGOVAY        TREMP FSMTKEESKPVQMMCMNNSFNVATL   PAEKMKILELPYASGDLSMLVLLPDE    VSGLERIEKTINFDKLREWTS TNAMAKKSMKVYLPRMKIE  EKYNLTSILMALG
HSAPLA        TQRDS FHLDEQFTVPVEMMQARTYPLRWFL  LEQPEIQVADFPFKN NMSFVVLVPTH    FEWNVSQVLANLSWDTLHPPL     VWERPTKVRLPKLYLK  HQMDLVATLSQLG
HSC1INHB      TRMEP FHF KNSVIKVPMMNSKKYPVAHFI  DQTLKAKVGQLQLSH NLSLVILVPQN    LKHRLEDMEQALSPSVFKAIMEKLEMSKFQPTLLTLPRIKVT  TSQDMLSIMEKLE
MSPROI        TMDRD FHVSKDKTIKVPTMIGKKDVRYADV  PELDAKMIEMSYEGD QASMIIILPNQ    VDGITALEQKLKDPKALSRAEERLYNTE  VEIYLPKFKIE  TTTDLKEVLSNMN
M24217        TYTDK FYISKNIVTSVDMMVSTENNLQYVHINELFGGFSIIDIPYEG NSSMVIILPDD    IEGIYNIEKNITDEKFKKWC  GMLSTKSIDLYMPKFKVEMTEPYNLVPILENLG
M24218        TSDYP FYVSPTEMVDVCIMSMYGEAFNHASVTESFGNFSIIELPYVG DTSMVVILPDN    IDGLESIEQNLTDTNFKKWC  NSMEATFIDVHIPKFKVT  GSYNLVDTLVKSG
PXCPVWPV      TSDYP FYVSPTEMVDVSMMSMYGEAFNHASVKESFGNFSIIELPYVG DTSMVVILPDN    IDGLESIEQNLTDTNFKKWC  DSMDAMFIDVHIPKFKVT  GSYNLVDALVKLG
PXVACB01      TSDYP FYVSPTEMVDVSMMSMYGKAFNHASVKESFGNFSIIELPYVG DTSMMVILPDK    IDGLESIEQNLTDTNFKKWC  NSLEATFIDVHIPKFKVT  GSYNLVDTLVKSG
OCRFVHOM C    TTDQP FY SGNVTYKVRMMNKIDTLKTETF TLRNVGYSVTELPYKRRQTAMLLVVPDD     LGEIVRALDLSLVRFWI  RNMRKDVCQVVMPKFSVE  SVLDLRDALQRLG
HSGDN         TKKRT FVAADGKSYQVPMLAQLSVFRCGSTSAPNDLWYNFIELPYHGESISMLIALPTE  SSTPLSAIIPHISTKTIDSWM  SIMVPKRVQVILPKFTAV  AQTDLKEPLKVLG
RRGDN         TKKRT FVAGDGKSYQVPMLAQLSVFRSGSTKTPNGLWYNFIELPYHGESISMLIALPTE  SSTPLSAIIPHISTKTINSWM  NTMVPKRMQLVLPKFTAL  AQTDLKEPLKALG
BTPAI1MR      THHRL FHKSDGSTISVPMMAQTNKFNYTEFTTPDGRYYDILELPYHGNTLSMLIAAPYE  KEVPLSALTSILDAELISQWK  GNMTRLTRLLVLPKFSLE  TEIDLRRPLENLG
HSPAIR        THRRL FHKSDGSTVSVPMMAQTNKFNYTEFTTPDGHYYDILELPYHGDTLSMFIAAPYE  KEVPLSALTNILSAQLISHWK  GNMTRLPRLLVLPKFSLE  TEVDLRKPLENLG
OAUMPA        TQKED FFLNDKTKVQVDMMRKTEQMLYSRS   EELFATMVKMPFKG NVSLILMLPDA    GHFDNALKKLTAKRAKLQK  ISNFRL VHLTLPKFKIT  FDINFKHLLPKIN
SSUFBP        TKKED FFVNEKTIVQVDMMRKTERMIYSRS   EELLATMVKIPCKE NASIILVLPDT    GKFNFALKEMAAKRARLQK  TNDFRL VHLVVPKIKDN  LQDRFKHLLPKIG
Phylip        +++++ ++++++++++++++++++++++++    ++++++++++++++ ++++++++++    ++++++++++++++++++    ++++++++++++++++  ++++++++++++
AT Structure  hF1-- s4C--||---s3C----------  ^hF2^----s2B--  ---s3B---     ^^hG^^  ^^^^^hH^^  ^^    --s2C-||-s6A  ----^^^hI^^^
AT Numbering        210       220       230       240       250       260       270       280       290       300
```

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES


```
                481       491       501       511       521       531       541       551       561       571       581       591
AT Numbering     310       320         3        30       340       350       360       370       380       390
AT Structure     ^hI1                ----      ---s5A------||--s4A-------- > <  --      s1C- --s4B--^hI2^---s5B---^hI3
RNSEPI1       IKKVFS QQADLSRITGT  KDLYV       SQVVHKAVLDVDETGTEATAATGVATVIR RQPRT     LNFNRPFMVVITDMDSQSILFVAKITNPK
RNSEPI2       IRKIFS QQADLSRITGT  KNLHV       SQVVHKAVLDVDETGTEGAAATAVTAALK SLPQT   IPPLNFNRPFMLVITDNNGQSVFFMGKVTNPM
RNSEPI3       IKEVFS TQADLSGITGD  KDLMV       FQVVHKAVLDVAETGTEAAAATGVKFVPM SAKLD   PLIIAFDRPFLMIISDTETAIAPFLAKIFNPK
RNSPI2        IRKIFS QQADLSRITGT  KNLHV       SQVVHKAVLDVDETGTEGAAATAVTAALK SLPQT   IPLLNFNRPFMLVITDNNGQSVFFMGKVTNPM+
RNSPI23       IRKIFS QQADLSRITGT  KNLHV       SQVVHKAVLDVDETGTEGAAATAVTAALK SLPQT   IPLLNFNRPFMLVITDNNGQSVFFMGKVTNPM
HSA1ACM       IEEAFT SKADLSGITGA  RNLAV       SQVVHKVVSDVFEEGTEASAATAVKITLL SALVE  TRTIVRFNRPFLMIIVPTDTQNIFFMSKVTNPSKPRACIKQWGSQ
PIGA1ACM      IQKIFS EEADLSGITDT  TPLCV       SQVVHSAVLDVDEEGTEGAASTGVVIERK SFENF     IVRFDSPFLFARSSETPRASSFWAKSSTPIQPRAPL
HSA1ATP       ITKVFS NGADLSGVTEE  APLKL       SKAVHKAVLTIDEKGTEAAGAMFLEAIPM SIPPE      VKFNKPFVFLMIEQNTKSPLFMGKVVNPTQK
PPATRP        ITKVFS NGADLSGVTED  APLKL       SKAVHKAVLTIDEKGTEAAGAMFLEAIPM SIPPE      VKFNKPFVFLMIEQNTKSPLFIGKVVNPTQK
MMAAT         ITRIFN NGADLSGITEEN APLKL       SKAVHKAVLTIDETGTEAAAATVFEAVPM SMPPI      LRFDHPFLFIIFEEHTQSPIFVGKVVDPTHK
OAPIA1T       INRVFS NGADLSGITEE  QPLMV       SKALHKAALTIDEKGTEAAGATFLEAIPM SLPPD      VEFNRPFLCILYDRNTKSPLFVGKVVNPTQA
HSCBG         IADLFT NQANFSRITQD  AQLKS       SKVVHKAVLQLNEEGVDTAGSTGVTLNLT SKPII      LRFNQPFIIMIFDHFTWSSLFLARVMNPV
HSCINHP       ISNVFT SHADLSRISNH  SNIQV       SEMVHKAVVEVDESGTRAAAATGTIFTFR SARLN   SQRLVFNRPFLMFIVD   NNILFLGKVNRP
HSTBG         IQHAYS ENADFSGLTED  NGLKL       SNAAHKAVLHIGEKGTEAAAVPEVELSDQ PENTF  LHPIIQIDRSFMLLILERSTRSILFLGKVVNPTEA
HSHCII        IRMLFD KNGNMAGISDQ  RIAID       LFKHQGTITVNEEGTQATTVTTVGFMPLS TQVRF      TVDRPFLFLIYEHRTSCLLFMGRVANPSRS
HSANG         LPAILH TELNLQKLSND  RIRVG       EVLNSIFFELEADEREPTESTQQLNKPEV LEVTL       NRPFLFAVYDQSATALHFLGRVANPLSTA
RNANG         LSTLLG AEANLGKMGDT  NPRVG       EVLNSILLELQAGEEEQPTESAQQPGSPE VLDVT       LSSPFLFAIYERDSGALHFLGRVDNPQNVV
HSATIII       LVDLFSPEKSKLPGIVAEGRDDLYV       SDAFHKAFLEVNEEGSEAAASTAVVIAGR SLNPN   RVTFKANRPFLVFIREVPLNTIIFMGRVANPCVK
HSPAI2        MEDAFNKGRANFSGMSER  NDLFL       SEVFHQAMVDVNEEGTEAAAGTGGVMTGR TGHGG    PQFVADHPFLFLIMHKITKCILFFGRFCSP+
MOUSE PAI2    MEDAFNKGKANFSGMSER  NDLFL       SEVFHQASVDVTEEGTVAAGGTGAVMTGR TGHGG    PQFVADHPFLFFIMDKITHTILFVGRFSSP
HVPROTZG      LQLPFS EEADLSEMVDSS QGLEI       SHVFHKSFVEVNEEGTEAGAATVAMGVAM SMPLK VDLVDFVANHPFLFLIREDIAGVVVFVGHVTNPLISA
CHKOVALM      ITDVFS SSANLSGISSA  ESLKI       SQAVHAAHAEINEAGREVVGSAEAGVDAA SVSEE      FRADHPFLFCIKHIATNAVLFFGRCVSP+
GGOVAY        MTDLFS RSANLTGISSV  DNLMI       SDAVHGVFMEVNEEGTEATGSTGAIGNIK HSLEL    EEFRADHPFLFFIRYNPTNAILFFGRYWSP
HSAPLA        LQELFQ  APDLRGISEQ   SLVV       SGVQHQSTLELSEVGVEAAAATSIAMSRM SLSSF       SVNRPFLFFIFEDTTGLPLFVGSVRNPNPSAPRELKEQQDSPGNK
HSC1INHB      FFD FS YDLNLCGLTED  PDLQV       SAMQHQTVLELTETGVEAAAASAISVART LLVFE       VQQPFLFVLWDQQHKFPVFMGRVYDPRA
MSPROI        IKKLFTPGAARLENLLKTK ESLYV       DAAIQKAFIEVNEEGAEAAAANAFGIVPA SLILY    PEVHIDRPFYFELK  IDGIPMFNGKVIEP
M24217        LTNIFG YYADFSKMCNETI   TV       EKFLHTTFIDVNEEYTEASAVTGVFMTNF SMVYR    TKVYINHPFMYMIKD NTGRILFIGKYCYPQ
M24218        LTEVFG STGDYSNMCNS  DV SV       DAMIHKTYIDVNEEYTEAAAATCALVSDC ASTIT    NEFCVDHPFIYVIRH VDGKILFVGRYCSPTTNC
PXCPVWPV      LTEVFG STGDYSNMCNS  DV SV       DAMIHKTYIDVNEEYTEAAAATCALVADC ASTVT    NEFCADHPFIYVIRH VDGKILFVGRYCSPTTN
PXVACB01      LTEVFG STGDYSNMCNS  DV SV       DAMIHKTYIDVNEEYTEAAAATCALVSDC ASTIT    NEFCVDHPFIYVIRH VDGKILFVGRYCSPTTNC
OCRFVHOM C    VRDAFDPSRADFGQASPS  NDLYV       TKVLQTSKIEADERGTTASSDTAITLIPR NALTA      IVANKPFMFLIYHKPTTTVLFMGTITKGEKVIYDTEGRDDVVSSV
HSGDN         ITDMFDSSKANFAKITRS  ENLHV       SHILQKAKIEVSEDGTKASAATTAILIAR SSPPW      FIVDRPFLFFIRHNPTGAVLFMGQINKP
RRGDN         ITEMFEPSKANFAKITRS  ESLHV       SHILQKAKIEVSEDGTKAAVVTTAILIAR SSPPW      FIVDRPFLFCIRHNPTGAILFLGQVNKP
BTPAI1MR      MTDMFRPSQADFSSFSDQ  EFLYV       SQALQKVKIEVNESGTLASSSTALVVSAR MAPEE      IIMDRPFLFVVRHNPTGTVLFMGQVMEP
HSPAIR        MTDMFRQFQADFTSLSDQ  EPLHV       AQALQKVKIEVNESGTVASSSTAVIVSAR MAPEE      IIMDRPFLFVVRHNPTGTVLFMGQVMEP
OAUMPA        LKHLLPKIDPKHTLTTTA  SSQHVTLKAPLPNLEALHQVEIELSEHALTTDTAIHTDNLLK VPANTKEVPVVVKFNRPFLLFVEDEITQTDLFVGQVLNPQVE
SSUFBP        INDIFT TKAVTWNTTGT  STI         LEAVHHAVIEVKEDGLTKNAAKDKDF WK VPVDKKEVPVVVKFDRPFFLFVEDEITRRDLFVAKVFNPKTE
Phylip        ++++++ ++++++++++++ +++++       ++++++++++++++++++++++++++++ +++++       ++++++++++++++++++++++++
AT Structure   ^hI1                ----      ---s5A------||--s4A-------- > <  --      s1C- --s4B--^hI2^---s5B---^hI3
AT Numbering     310       320         3        30       340       350       360       370       380       390

                601       611       621       631       641       651       661       671       681       691       701       711
AT Numbering
AT Structure
HSAPLA        DFLQSLKGFPRGDKLFGPDLKLVPPMEEDYPQFGSPK
Phylip
AT Structure
AT Numbering
```

range, a phenomenon caused by the presence of a number of highly related sequences in the database.

Examination of the sequence alignment reveals a number of features. The A helix is variable in length being shortened by three or four residues, or one turn at the amino terminal end, in the ovalbumin/PAI2 families and truncated by almost half in the pox virus serpins. This suggests that this helix is not crucial for structure despite its prominent position in ovalbumin and antitrypsin. In antithrombin the region corresponding to the A-helix is preceded by about 45 residues which play a significant role in the binding of heparin to antithrombin and appear to be involved in heparin-binding. Longer extensions are found in heparin cofactor II, antiplasmin, C_1-inhibitor and the angiotensinogens. It is not clear that these regions form the same compact fold that is assumed for the antithrombin extension. In C_1-inhibitor at least, it appears that this region forms an extended, highly glycosylated 'brush' that is distinct from the serpin fold. The angiotensinogens are a special case as the amino-terminal region is the origin for angiotensin which is released from the rest of the molecule in a set of specific cleavages, and it is likely that this region also forms an extended structure that allows proteases access to the cleavage sites.

A block of strongly conserved residues are found in strand six of sheet B (s6B) and helix B, including an almost completely conserved Ser–Pro dipeptide at antitrypsin residues 53 and 54. The region between helix C and helix D, including helix C1, is very variable in length and includes an insertion of about 30 residues in PAI2 sequences. In addition, many of the helix D residues are absent in the pox virus serpins, and this helix is completely absent in the malignant rabbit fibroma virus serpin (OCRFVHOM C). Residues in the D-helix have been implicated in heparin-binding in several serpins, namely heparin cofactor II, antithrombin and protease nexin (Craig *et al.* 1989; Sun & Chang 1989*a*; Sun & Chang 1989*b*; Borg *et al.* 1990; Evans *et al.* 1990; Loganathan *et al.* 1990; Whinna *et al.* 1991; Evans *et al.* 1992), and it is possible that this region, which forms a 'face' is important in ligand binding and so might be expected to be variable.

Further conserved regions of sequence are present in sheet 2A and helix E. Helix F sits above and approximately parallel with the A β-sheet. The peptide strand forms a bend after the F helix and returns alongside it in a series of β-bulges (Loebermann *et al.* 1984). Within the F helix occurs a highly conserved dipeptide Ile-Asn, where a hydrogen bond between the sidechain of the asparagine residue and the backbone of the β-bulges serves to pin these structures together. This hydrogen bond would appear to be important in the conservation of the structure of the F-helix and bulges. Further conserved regions are found in many of the regions of secondary structure, particularly strands 3A, 4C and 3C, where it is likely they play some role in maintaining those structures.

At the boundary between strands 5A and 4A lies an area of strong sequence conservation, particularly a glutamine at antitrypsin 342. A mutation at this Glu to Lys in antitrypsin causes the Z antitrypsin phenotype associated with the formation of antitrypsin aggregates in the liver of affected individuals. Recent work has shown that this mutation favours the insertion of s4A into the A-sheet either by releasing strand 4 from its correct fold, or by making the A sheet more receptive to the insertion of the s4A loop (Lomas *et al.* 1992). The consequence of this interaction is the formation of extensive aggregates with serious effects on the liver of affected individuals. This Glu is conserved in all but two of the serpins shown in figure 2 – human and rat angiotensinogen – and these two serpins are not known to undergo the characteristic change of thermal stability after cleavage and are not known to be inhibitory.

The region of the reactive centre loop, s4A in antitrypsin, is quite variable in composition particularly in the latter part of the strand, as are the P_1 and P_1' residues that form the active site itself. The region between Glu342 and the reactive centre at P_1–P_1' is quite conserved at the amino terminal end, a consequence probably of the specific requirements for the reinsertion of the strand into the sheet (Huber &

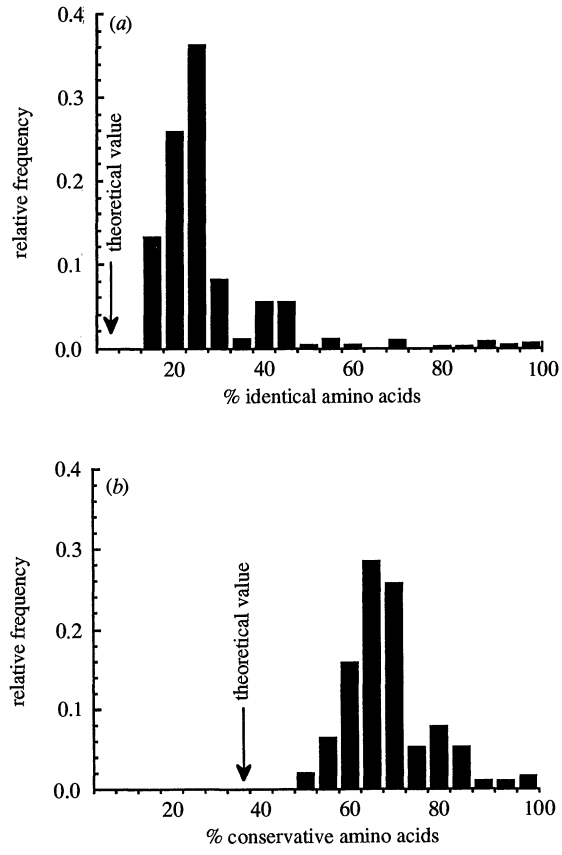

Figure 3. Distribution of homology scores. In each case the scores have been divided into bins of width 5% and the histogram plotted. (*a*) The scores obtained by calculation of pairwise percentage identities are compared with the value expected for a pair of sequences of the same composition as the serpins but of random sequence order. (*b*) Pairwise conservative amino acid replacement scores are shown compared with the expected conservative value of a pair of sequences of random order but of a composition similar to that of the serpins.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Table 2. Differences among four serpins superimposed either by a sequence or a structural basis. Structures were superimposed and the root mean square (RMS) differences in Å between equivalent C_α atoms calculated and shown for both sequence and structure based superpositions, and for the difference between them.

molecules compared	RMS differences in angstroms (Å)		
	sequence	structure	difference
antichymotrypsin			
vs antitrypsin	0.709	0.659	0.050
vs ovalbumin	1.534	1.468	0.066
vs PAI1	1.740	1.676	0.064
antitrypsin			
vs ovalbumin	1.922	1.596	0.326
vs PAI1	1.631	1.594	0.037
ovalbumin			
vs PAI1	1.723	1.708	0.015

Carrell 1989; Carrell *et al.* 1991), but becomes more variable closer to the reactive centre itself. The variability here probably reflects different specificity amongst the serpins and the differing requirements for interactions with the appropriate cognate protease.

Strand 1 of sheet C (s1C) is also quite variable in length and composition although at the boundary between it and s4B there is a conserved region including a Pro–Phe pair. It has been suggested that this region forms a 'stalk' important both for function and perhaps for control of specificity (Huber & Carrell 1989; Engh *et al.* 1990; Mottonen *et al.* 1992). The region on this side of the reactive centre seems to act as an anchor for the retention of structure after cleavage and subsequent final folding and displays considerable regions of local sequence homology among all the serpins.

A more complete account of structurally significant residues and their role in serpin architecture can be found in Huber & Carrell (1989).

(*b*) *Structural analysis*

To make an estimate of the reliability of the alignment shown in figure 2, superpositions were made of the four available serpin structures. Superpositions were based either on the sequence alignment where the three-dimensional superposition was based around amino acids considered to be related in the sequence alignment, or on secondary structure elements common to both structures. Table 2 shows the result of these superpositions, where both the sequence and structure based figures are presented, as is the difference reported between each method. Superposition based on common structural units is to be preferred where both structures are available as the actual positions of amino acids in both structures are known. The more accurate a predictor of structure is the sequence alignment, the smaller should be the variations in the fit of the different superpositions based either on sequence or structure. The data presented in figure 2 suggest that this was so. The variations in fit are from 0.015 Å for ovalbumin–PAI1, 0.064 Å for antitrypsin–antichymotrypsin, to 0.326 Å for the ovalbumin–antitrypsin comparison. Examination of the data in table 2 shows that that latter figure is the most different. Examination of the superimposed structures showed that the region of poor fit was in the helix C, C1 and D region where the greatest diversity of length in the sequences occurs. About five residues are poorly aligned here and account for most of the variation observed in the antitrypsin–ovalbumin pairing. It is noteworthy that the prediction for antichymotrypsin–ovalbumin comparison is not similarly incorrect. That these variations are modest suggests that the alignment predicts the secondary structure reasonably well.

A distinction must be drawn between these variations and the actual root mean square difference. The root mean square difference would be expected to grow as the sequence similarity declined, until at values of sequence identity of only 10–15%, the root mean square deviation reaches about 2.4 Å (Chothia & Lesk 1986) and seems to be related to the finding that structural elements remain well-conserved but the exact orientation of the elements changes (Lesk & Chothia 1980, 1982; Chothia & Lesk 1986; Lesk *et al.* 1986; Chothia & Finkelstein 1990). Although only four serpins were able to be analysed in this way, the results suggest some confidence may be placed in the sequence alignment. The accuracy of the sequence alignment is essential to the determination of the phylogenetic trees as even a small area of misplaced sequence may lead to quite incorrect trees.

(*c*) *Gene tree analysis*

The best method for calculating gene trees from sequence data is an issue that is still much debated. Much of this debate centres on whether distance or parsimony methods are more appropriate for phylogenetic analysis of sequence data. Recent work has suggested that for a completely known phylogeny of strains of phage T7, maximum parsimony analysis gave the answer most consistent with the known phylogeny (Cunningham *et al.* 1992). A number of methods of phylogenetic analysis were considered including heuristic methods such as maximum likelihood (Felsenstein 1981, 1988) but preliminary work indicated these methods gave inconsistent and contradictory results. For this analysis the maximum parsimony method was adopted which gave rise to some problems. The number of possible unrooted trees N for n sequences (or different characters) is given by

$$N=\frac{(2n-5)!}{2^{n-3}(n-3)!}, \qquad (3)$$

and so to determine the minimum length tree by examining every possible arrangement of the branches for the thirty seven sequences analysed here would require the examination of about 3.4×10^{49} arrangements. Since only some 10^{18} seconds have elapsed since the beginning of the Universe, this is clearly an

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

impracticable number of trees to analyse and this is a member of the class of NP-complete problems for which no efficient solution exists. Even when using the branch-and-bound algorithm of Hendy & Penny (1982) to simplify the solution, the problem remains intractable. Consideration was given to reducing the size of the data set by removing the third codon position. However many informative sites would have been lost by this procedure and although the length of the alignment would have been reduced by a third, the number of sequences would have remained unaltered and the number of possible trees remained the same. In general, removal of the third codon is a useful procedure for datasets where the GC content of the sequences varies widely, but has little advantage in other circumstances.

The 'bootstrap' method was employed (Felsenstein 1985, 1988) to sample the possible tree-space and to gain information about the reliability of the trees. In this method, the same set of species is varied by duplicating some characters and dropping others, to leave the same number of sites as in the original data. By doing this, a distribution is created comparable to the sample of unknown distribution from which the original data were drawn. Subsequent analysis of the trees produced, and the calculation of the most commonly occurring tree or trees, allows a statistical estimate of the reliability of each branch. This procedure is not guaranteed to find either the most parsimonious or the best tree, but provides a good estimate of the actual phylogenetic tree.

Sequence alignment for tree analysis involved the removal of regions of the alignment where there were significant insertions or deletions. The rationale for this pruning was two-fold. First, the difficulty of ascribing a reasonable penalty value to such regions. For example, in human and mouse plasminogen activator inhibitor 2 (HSPAI2 and MUSPAI2) there is an insertion of 25 amino acids with respect to ovalbumin and more with respect to many of the other sequences. It is difficult to ascribe a score to such an event; should it be considered as one mutation, as 25 individual changes, or as some intermediate value? The significant variation in the length of this region suggests that different evolutionary constraints operate in the different molecules. Indeed, modelling of this region in human PAI2 suggests an added loop packing against the A-helix where there are a number of mutations from a charged to a neutral amino acid (C. Marshall, unpublished data).

Second, there is also the question of what role highly variable areas may have in the protein. Regions such as the insertion in PAI2 sequences may confer unique properties on each serpin, but may not reveal much about common features since the structural and functional, and hence evolutionary constraints, may be quite different. For these reasons, regions at the amino and carboxyl termini, unique to specific serpins, were removed from consideration. The signal sequences, where applicable (and it should be noted that PAI2 and ovalbumin lack cleavable amino-terminal signal peptides and have instead an internal signal sequence (Palmiter *et al.* 1978; Ye *et al.* 1988, 1989; von Heijne *et al.* 1991)) were also removed as it is clear that homology in these regions reflects functions associated with export from the cell and does not contribute to the serpin architecture.

The most parsimonious trees were then calculated using a limited rearrangement of nodes. All the trees so produced were analysed to find a majority-rule consensus tree; that is, all the groups that occur in more than half of the trees produced by the bootstrap estimation. This was further refined by including those groups which further resolved the tree and which did not contradict more frequently occurring groups. Also calculated was the relative frequency with which each branch appeared in the final consensus tree. These data are shown in figure 4 and can be seen to range from about 8% to 100%, although higher values predominate. The higher each value is, the more reliable that particular branch is in the overall alignment. It is clear that outer branches consistently score higher values than some of those closer to the 'root' of the tree. Overall the tree carries a low statistical significance, although many of the branches are well-supported, and consideration of the tree must be made with this caveat in mind.

Phylogenetic relationships amongst the serpins calculated by the procedure above are shown in figures 4 and 5. The gene tree in figure 5 better shows the relationships among the whole superfamily of sequences, whereas the cladogram in figure 4 indicates the connections within the groups more clearly. The gene tree shows the presence of three main branches. The first of these branches (figure 5) contains the outgroup barley protein Z (HVPROTZG), the insect serpin from *Manduca sexta*, the ovalbumin and PAI2 family, the pox virus serpins (M24217, M24218, PXVAXB01 and PXCPVWPV), and antithrombin (HSATIII). This branch of serpins is the most diverse and contains almost all the non-mammalian serpins. Scores in this part of the tree are reasonably high suggesting some reliability for this branch. The value of 50% determined in the arrangement of vaccinia inhibitor 2, vaccinia B13R and cowpox CPV-W2 serpins, reflects a trivial rearrangement of these three sequences. Less reliability is found for the antithrombin branch, although examination of many of the trees showed this branch to be in this region, if not in precisely this position.

The grouping of PAI2 and the ovalbumin related sequences is well supported and is perhaps surprising considering that the divergence of the birds and mammals is thought to have occurred about 180 million years ago. The finding that these sequences lack an amino terminal signal sequence, having instead an internal sequence that serves the same purpose (Palmiter *et al.* 1978; Ye *et al.* 1988, 1989; von Heijne *et al.* 1991), gives support to their being related. It is likely that ovalbumin has changed from being inhibitory to acting as storage protein. The role of the closely related ovalbumin gene Y (and the partly sequenced gene X protein) is not known, but is likely to be similar to that of the ancestral ovalbumin/PAI2 protein.

The second main branch contains the antitrypsin,

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

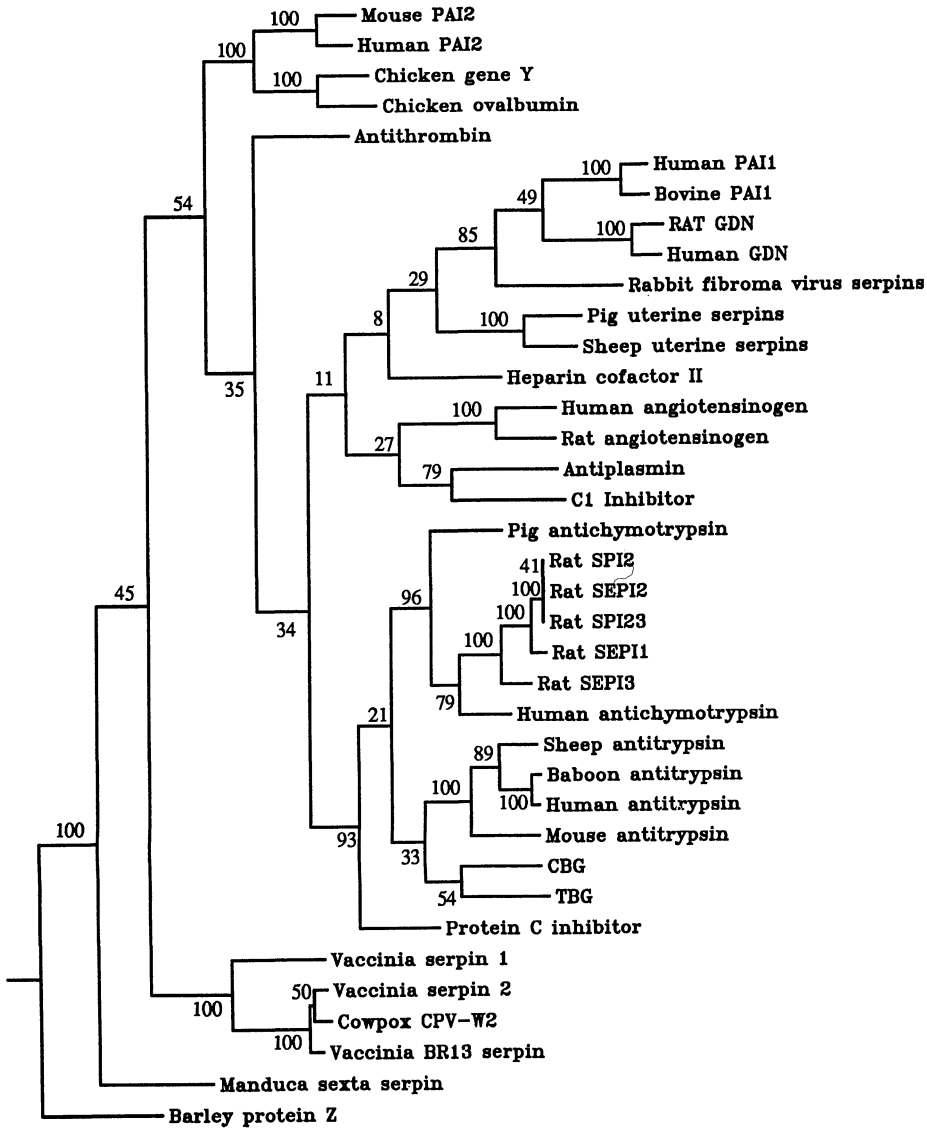

Figure 4. Gene tree of the serpins calculated as indicated in the text. This representation shows the relationship amongst branches of the tree. Note that the sequence HVPROTZG (barley protein Z) was defined as the outgroup as *a priori* it was considered the most evolutionary distant sequence. Branch lengths are proportional to evolutionary distance. All branch tips, unless stated otherwise, are human.

antichymotrypsins and the related sequences. This overall branch has a high percentage occurrence although some of the sub-branches are less common. Cortisol and thyroxine binding globulins (CBG and TBG) are suggested to be closely related and to be members of the antitrypsin family. Variation in the branch placing here consisted mostly of CBG being placed as a branch of the antitrypsin grouping and the TBG branch ancestral to that. The tree for the antitrypsin is sufficiently strong to be authoritative and shows the branching pattern expected from conventional phylogenetic data. The rat SPI and SEPI sequences appear to be members of the antichymotrypsin branch, and the 41% at the branch involving rat SP12, SP123 and SEP12 reflects a trivial rearrangement of the very closely related branches. The branching pattern suggests that human and pig antichymotrypsins are not particularly closely related, and these, and the last member of this branch, human protein C inhibitor (HSCINHP) receive strong support from the fraction of times that branch occurs. The support for the relationship between the antitrypsin and antichymotrypsin branches is low, although examination of alternative trees suggests the basic pattern is correct.

The last of the main branches is less reliable than the other two. Many of the branches occur infrequently, and this should be considered the most speculative branch of the whole tree. The angiotensinogens (HSANG and RNANG) appear to be related to antiplasmin (HSAPLA) and C_1-inhibitor (HSC1INHB), either from a single common ancestor, or having diverged from the parent branch one after the other. All these sequences have amino terminal extensions. In the angiotensinogens this region is processed to form angiotensin, a peptide active in control of blood pressure. C_1-inhibitor, as discussed above, has an extra, highly glycosylated domain at the amino terminus. Antiplasmin also has an amino terminal extension although no specific function has

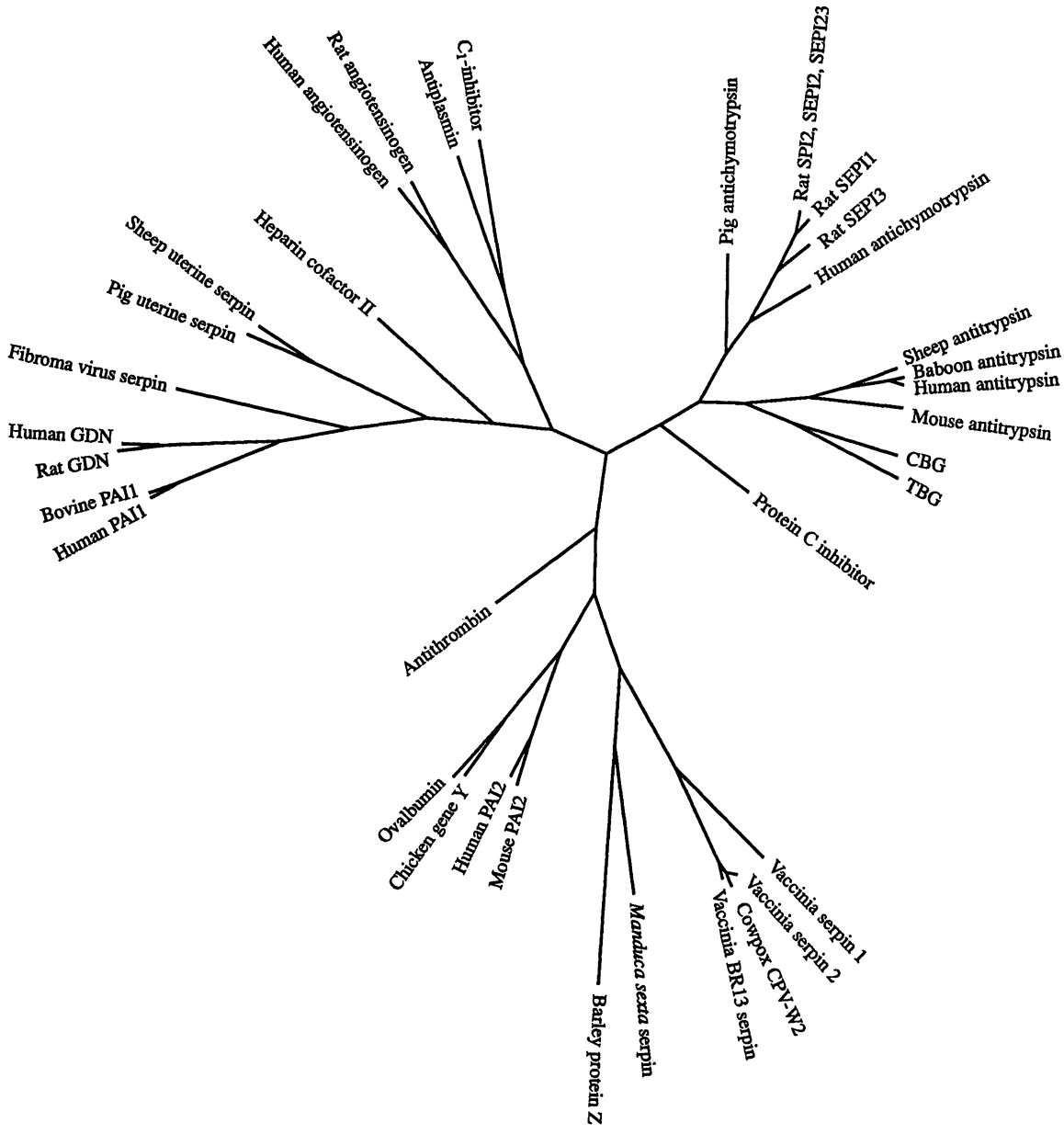

Figure 5. Cladogram of the serpins calculated as indicated in the text. This representation shows the relationship among close relatives of the tree. Note that the sequence HVPROTZG (barley protein Z) was defined as the outgroup as *a priori* it was considered the most evolutionarily distant sequence. Branch lengths are proportional to evolutionary distance. All branch tips, unless stated otherwise, are human. The numbers at each branch are, as a percentage, the fraction of times a particular branch occurs in all the trees generated by the SEQBOOT and DNAPARS programs.

been ascribed to it. These data provide some biological evidence of relatedness.

Heparin cofactor II (HSHCII) is a member of the third main branch. Physiologically heparin cofactor II appears to be related in function to antithrombin and the control of blood-clotting, and both are activated by sulfated polysaccharides such as heparin. However the phylogenetic evidence suggests that the sequences are not particularly closely related and their similarity of function is more likely to be the consequence of convergent evolution. The exact placement of the heparin cofactor II branch is very poorly supported, the branch shown here being present in only 8% of the total trees. However, in very few of those alternative trees were antithrombin and heparin cofactor II considered to be related.

Also on this branch are sheep uterine milk protein (OAUMPA) and pig uteroferrin associated protein (SSUFBP). The function of these serpins is unknown but they appear to play some role in reproduction. A serpin (OCRFVHOM C) initially isolated from a rabbit cell line, and later shown to be derived from a fibroma virus infecting the cells (Upton *et al.* 1986), is related to this branch. It is likely that this sequence is derived from a rabbit serpin in this group. Furthermore, this viral serpin appears to be rather distantly related to the pox virus serpins (M24217, M24218, PXVAXB01 and PXCPVWPV).

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

The final grouping on this branch is that of the glial-derived nexins (HSGDN and RRGDN) and the plasminogen activator inhibitors I (BTPAIMR and HSPAIR). These serpins are both involved in tissue remodelling and appear to be relatively closely related. The relatively low proportion of times this branch occurs (49%) is partially related to variation in the placement of the rabbit fibroma virus branch rather than placement of either the PAI1 or glial-derived nexin (GDN) branches elsewhere.

4. DISCUSSION

The serpins are a family of proteins widely distributed throughout the animal and plant kingdom and possessing a diverse range of functions. Despite the evolutionary distances among serpins and their different functions, the overall serpin structure appears highly conserved. The conservation of structure suggests that the serpin architecture can tolerate only a limited range of changes. Evidence for this is particularly apparent when the structures of the human proteinase inhibitor antitrypsin and the chicken egg storage protein ovalbumin are compared. It has been suggested that intact serpins can be considered as trapped folding intermediates, which can only reach a final, stable state after cleavage (Loebermann *et al.* 1984). Some physical evidence favouring this view has been adduced which suggests that intact and functional serpins have strained secondary structural elements that adopt a relaxed form upon cleavage and rearrangement (Gettins & Harten 1988; Gettins 1989; Haris *et al.* 1990; Carrell *et al.* 1991). It is possible that it is these constraints that have maintained the similarity of structures and sequences within the serpin family.

The alignment described in this paper has been made by considering both sequence data and structural information. A number of features appear that assist in aligning new sequences. Confirmation of the overall correctness of the alignment comes from the similarity of structural superpositions based either on sequence or secondary structure data. This allows the alignment to be used with some confidence in identifying regions of interest in new serpins, and should allow the more reliable identification of the active site and other regions of significance in new serpin sequences.

The use of the bootstrap in conjunction with parsimony analysis has been found to be a powerful technique and has been widely used in identifying plausible phylogenetic trees (Thomas *et al.* 1989; Tristem *et al.* 1990, 1992; Cunningham *et al.* 1992; Hillis *et al.* 1992). The danger of not adequately exploring the possible trees inherent in a dataset has been emphasized by recent work questioning the phylogenetic analysis of a human mitochondrial dataset by Wilson (Cann *et al.* 1987; Vigilant *et al.* 1991; Gibbons 1992; Thorne & Wolpoff 1992; Wilson & Cann 1992).

A striking finding of the serpin gene tree and cladogram is the evolutionary diversity among mammalian serpins when compared with those serpins expected to be evolutionarily distant. For example, the evolutionary distance found in this analysis between human antitrypsin and glial-derived nexin is of a similar magnitude to that between human antitrypsin and barley protein Z. Furthermore, the tree distance between barley protein Z and the serpin from the insect *Manduca sexta* is relatively small despite the large evolutionary distance between these organisms. Even with constraints limiting the degree to which serpins can mutate and remain serpins, it is surprising that the tree distance is not greater; perhaps of a similar size to the distances found among the mammalian serpins. It is possible that these relatively small evolutionary distances reflect a limitation on the amounts serpins can vary and remain serpins; a suggestion which may possibly relate to the strong architectural constraints on serpins discussed above. However, the variation among mammalian serpins remains surprisingly high. It is possible that there has been a significant radiation of serpin function in mammals, and possibly in birds (although it is possible that this observation comes from sampling biased heavily toward mammalian serpins).

To examine this possibility it would be instructive to look for other serpins in birds and possibly in reptiles and calculate their relationship with the mammalian serpins. Recent findings include novel serpins from wallabies (Patterson *et al.* 1991), a serpin found in the endoplasmic reticulum of myoblasts (Clarke *et al.* 1991), and another associated with kallikrein-binding (Chai *et al.* 1991). An example of an important area where serpins have newly been found to have a significant role is in the control of neural growth and remodelling where both PAI1 and GDN have been shown to be important (Monard *et al.* 1983; Reinhard *et al.* 1988; Wagner *et al.* 1989; Cunningham *et al.* 1990; Festoff *et al.* 1990; McGrogan *et al.* 1990; Seeds *et al.* 1990). If it is true that serpin function has expanded in mammals, then it is likely that many more serpins with new functions remain to be found in this group.

I acknowledge the valuable help and advice of Arthur Lesk, Peter Stockwell and John Cutfield. Rick Engh kindly provided antichymotrypsin coordinates after Baumann *et al.* (1991) and Betsy Goldsmith the coordinates of PAI1 after Mottonen *et al.* (1992). Joe Felsenstein provided advice and reassurance about producing phylogenetic trees. The author was the recipient of an HRC-Wellcome Overseas Research Fellowship and thanks Robin Carrell for the time spent in his laboratory.

REFERENCES

Abola, E.E., Bernstein, F.C., Bryant, S.H. & Koetzle, T.F. 1987 Protein data bank. In *Crystallographic databases – information content, software systems, scientific applications* (ed. F. H. Allen, G. Bergerhoff & R. Sievers), pp. 107–132. Bonn, Cambridge, Chester: International Union of Crystallography.

Barton, G.J. & Sternberg, M.J.E. 1987 A strategy for the rapid multiple alignment of protein sequences. *J. molec. Biol.* **198**, 327–337.

Baumann, U., Huber, R., Bode, W., Grosse, D., Lesjak, M. & Laurell, C.B. 1991 Crystal structure of cleaved human

α_1-antichymotrypsin at 2.7 Å resolution and its comparison with other serpins. *J. molec. Biol.* **218**, 595–606.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F. Jr, Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The protein data bank: a computer-based archival file for macromolecular structures. *J. molec. Biol.* **112**, 535–542.

Bode, W., Mayr, I., Baumann, U., Huber, R., Stone, S.R. & Hofsteenge, J. 1989 The refined 1.9 Å crystal structure of human α-thrombin: interaction with D-Phe-Pro-Arg chloromethylketone and the significance of the Tyr-Pro-Pro-Trp insertion segment. *EMBO J.* **8**, 3467–3475.

Borg, J.-Y., Brennan, S.O., Carrell, R.W., George, P., Perry, D.J. & Shaw, J. 1990 Antithrombin Rouen-IV 24 Arg→Cys. *FEBS Lett.* **266**, 163–166.

Brandt, A., Svendsen, I. & Hejgaard, J. 1990 A plant serpin gene; structure, organization, and expression of the gene coding barley protein Z_4. *Eur. J. Biochem.* **194**, 499–505.

Cann, R.L., Stoneking, M. & Wilson, A.C. 1987 Mitochondrial DNA and human evolution. *Nature, Lond.* **325**, 31–36.

Carrell, R.W., Evans, D.L. & Stein, P.E. 1991 Mobile reactive centre of serpins and the control of thrombosis. *Nature, Lond.* **353**, 576–578.

Chai, K.X., Ma, J.-X., Murray, S.R., Chao, J. & Chao, L. 1991 Molecular cloning and analysis of the rat kallikrein-binding protein gene. *J. biol. Chem.* **266**, 16 029–16 036.

Chothia, C. & Finkelstein, A.V. 1990 The classification and origins of protein folding patterns. *A. Rev. Biochem.* **59**, 1007–1039.

Chothia, C. & Lesk, A.M. 1986 The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826.

Clarke, E.P., Cates, G.A., Ball, E.H. & Sanwal, B.D. 1991 A collagen-binding protein in the endoplasmic reticulum of myoblasts exhibits relationship with serine protease inhibitors. *J. biol. Chem.* **266**, 17 230–17 235.

Craig, P.A., Olson, S.T. & Shore, J.D. 1989 Transient kinetics of heparin-catalyzed protease inactivation by antithrombin III. *J. biol. Chem.* **264**, 5452–5461.

Cunningham, C.W., Blackstone, N.W. & Buss, L.W. 1992 Evolution of king crabs from hermit crab ancestors. *Nature, Lond.* **355**, 539–542.

Cunningham, D.D., Farrell, D.H. & Wagner, S.L. 1990 Regulation of protease nexin I activity and target protease specificity by the extracellular matrix. In *Serine proteases and their serpin inhibitors in the nervous system* (ed. B. Festoff), pp. 93–102. New York and London: Plenum Press.

Engh, R.A., Wright, H.T. & Huber, R. 1990 Modelling the intact form of the α_1-proteinase inhibitor. *Prot. Engineer.* **3**, 469–477.

Evans, D.L., Christey, P.B. & Carrell, R.W. 1990 The heparin binding site and activation of protease nexin I. In *Serine proteases and their serpin inhibitors in the nervous system* (ed. B. Festoff), pp. 69–77. New York and London: Plenum Press.

Evans, D.L., Marshall, C.J., Christey, P.B. & Carrell, R.W. 1992 Heparin binding site, conformational change and activation of antithrombin. *Biochemistry* **31**, 12 629–12 642.

Felsenstein, J. 1981 A likelihood approach to character weighting and what it tells us about parsimony and compatability. *Biol. J. Linn. Soc.* **16**, 183–196.

Felsenstein, J. 1983 Parsimony in systematics: biological and statistical issues. *A. Rev. Ecol. System.* **14**, 313–333.

Felsenstein, J. 1985 Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783–791.

Felsenstein, J. 1988 Phylogenies from molecular sequences: inference and reliability. *A. Rev. Genet.* **22**, 521–565.

Festoff, B.W., Rao, J.S., Reddy, B.R. & Hantaï, D. 1990 A cascade approach to synapse formation based on thrombogenic and fibrinolytic models. In *Serine proteases and their serpin inhibitors in the nervous system* (ed. B. W. Festoff), pp. 245–253. New York and London: Plenum Press.

Gettins, P. 1989 Absence of large-scale conformational change upon limited proteolysis of ovalbumin, the prototypic serpin. *J. biol. Chem.* **354**, 3781–3785.

Gettins, P. & Harten, B. 1988 Properties of thrombin- and elastase-modified human antithrombin III. *Biochemistry* **27**, 3634–3639.

Gibbons, A. 1992 Mitochondrial Eve: wounded but not yet dead. *Science, Wash.* **257**, 873–875.

Haris, P.I., Chapman, D.C., Harrison, R.A., Smith, K.F. & Perkins, S.J. 1990 Conformational transition between native and reactive center cleaved forms of α_1-antitrypsin by fourier transform infrared spectroscopy and small-angle neutron scattering. *Biochem. J.* **267**, 203–212.

Hendy, M.D. & Penny, D. 1982 Branch and bound algorithms to determine minimal evolutionary trees. *Math. Biosci.* **59**, 277–290.

Hillis, D.M., Bull, J.J., White, M.E., Badgett, M.R. & Molineux, I.J. 1992 Experimental phylogenetics: generation of known phylogeny. *Science, Wash.* **255**, 589–592.

Huber, R. & Carrell, R.W. 1989 Implications of the three-dimensional structure of α_1-antitrypsin for structure and function of serpins. *Biochemistry* **28**, 8951–8966.

Hunt, L.T. & Dayhoff, M.O. 1980 A surprising new protein superfamily containing ovalbumin, antithrombin III, and alpha$_1$-proteinase inhibitor. *Biochem. biophys. Res. Commun.* **95**, 864–871.

Kanost, M.R., Prasad, S.V. & Wells, M.A. 1989 Primary structure of a member of the serpin superfamily of proteinase inhibitors from an insect. *Manduca sexta. J. biol. Chem.* **264**, 965–972.

Kraulis, P.J. 1991 MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. appl. Crystal.* **24**, 946–950.

Lesk, A.M. & Chothia, C. 1980 How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. molec. Biol.* **136**, 225–270.

Lesk, A.M. & Chothia, C. 1982 Evolution of proteins formed by β-sheet II: the core of the immunoglobin domains. *J. molec. Biol.* **160**, 325–342.

Lesk, A.M., Levitt, M. & Chothia, C. 1986 Alignment of amino acid sequences of distantly related proteins using variable gap penalties. *Prot. Engineer.* **1**, 77–78.

Loebermann, H., Tokuoka, R., Deisenhofer, J. & Huber, R. 1984 Human α_1-proteinase inhibitor. *J. molec. Biol.* **177**, 531–556.

Loganathan, D., Wang, H.M., Mallis, L.M. & Linhardt, R.J. 1990 Structural variation in the antithrombin III binding site region and its occurrence in heparin from different sources. *Biochemistry* **29**, 4362–4368.

Lomas, D.A., Evans, D.L., Finch, J.T. & Carrell, R.W. 1992 The mechanism of Z α_1-antitrypsin accumulation in the liver. *Nature, Lond.* **357**, 605–607.

McGrogan, M., Kennedy, J., Golini, F., Ashton, N., Dunn, F., Bell, K., Tate, E., Scott, R.W. & Simonsen, C.C. 1990 Structure of the human protease nexin I gene and expression of recombinant forms of PN-1. In *Serine proteases and their serpin inhibitors in the nervous system* (ed. B. W. Festoff), pp. 147–161. New York and London: Plenum Press.

McReynolds, L., O'Malley, B.W., Nisbet, A.D., Fothergill,

J.E., Givol, D., Fields, S., Robertson, M. & Brownlee, G.G. 1978 Sequence of chicken ovalbumin mRNA. *Nature, Lond.* **273**, 723–728.

Monard, D., Niday, E., Limat, A. & Solomonson, F. 1983 Inhibition of protease activity can lead to neurite extension in neuroblastoma cells. *Prog. Brain Res.* **56**, 359.

Mottonen, J., Strand, A., Symersky, J., Sweet, R.M., Danley, D.E., Geoghegan, K.F., Gerard, R.D. & Goldsmith, E.J. 1992 Structural basis of latency in plasminogen activator inhibitor-1. *Nature, Lond.* **355**, 270–273.

Needleman, S.B. & Wunsch, C.D. 1970 A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. molec. Biol.* **48**, 443–453.

Palmiter, R.D., Gagnon, J. & Walsh, K.A. 1978 Ovalbumin: a secreted protein without a transient hydrophobic leader sequence. *Proc. natn. Acad. Sci. U.S.A.* **75**, 94–98.

Patterson, S.D., Bell, K. & Shaw, D.C. 1991 The tammar wallaby major plasma serpins: partial characterization including the sequence of the reactive site region. *Comp. Biochem. Physiol.* **98C**, 359–367.

Reinhard, E., Meier, R., Halfter, W., Rovelli, G. & Monard, D. 1988 Detection of glia-derived nexin in the olfactory system of the rat. *Neuron* **1**, 387.

Seeds, N.W., Verral, S., McGuire, P. & Friedman, G. 1990 Plasminogen activator in the developing nervous system. In *Serine proteases and their serpin inhibitors in the nervous system* (ed. B. W. Festoff), pp. 173–184. New York and London: Plenum Press.

Stein, P.E., Leslie, A.G.W., Finch, J.T. & Carrell, R.W. 1991 Crystal structure of uncleaved ovalbumin at 1.95 Å resolution. *J. molec. Biol.* **221**, 941–959.

Stein, P.E., Leslie, A.G.W., Finch, J.T., Turnell, W.G., McLaughlin, P.J. & Carrell, R.W. 1990 Crystal structure of ovalbumin as a model for the reactive centre of the serpins. *Nature, Lond.* **347**, 99–102.

Stockwell, P.A. 1988 HOMED: a homologous sequence editor. *Trends Biochem. Sci.* **13**, 322–323.

Sun, X.-J. & Chang, J.-Y. 1989*a* The heparin and pentosan polysulfate binding sites of human antithrombin overlap but are not identical. *Eur. J. Biochem.* **185**, 225–230.

Sun, X.-J. & Chang, J.-Y. 1989*b* Heparin binding domain of human antithrombin III inferred from the sequential reduction of its three disulfide linkages. *J. biol. Chem.* **264**, 11 288–11 293.

Takagi, H., Naruma, H., Nakamura, K. & Sasaki, T. 1990 Amino acid sequence of silkworm *Bombyx mori* hemolymph antitrypsin deduced from its cDNA nucleotide sequence: confirmation of its homology with serpins. *J. biochem.* **108**, 372–378.

Thomas, R.H., Schaffner, W., Wilson, A.C. & Paäbo, S. 1989 DNA phylogeny of the extinct marsupial wolf. *Nature, Lond.* **340**, 465–467.

Thorne, A.G. & Wolpoff, M.H. 1992 The multiregional evolution of humans. *Scient. Am.* April 28–33.

Tristem, M., Marshall, C., Karpas, A. & Hill, F. 1992 Evolution of the primate lentiviruses: Evidence from vpx and vpr. *EMBO J.* **11**, 3405–3412.

Tristem, M., Marshall, C., Karpas, A., Petrik, J. & Hill, F. 1990 Origin of *vpx* in lentiviruses. *Nature, Lond.* **347**, 341–342.

Upton, C., Carrell, R.W. & McFadden, G. 1986 A novel member of the serpin super-family is encoded on a circular plasmid-like DNA species isolated from rabbit cells. *FEBS Lett.* **207**, 115–120.

Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K. & Wilson, A.C. 1991 African populations and the evolution of human mitochondrial DNA. *Science, Wash.* **253**, 1503–1507.

von Heijne, G., Liljestrom, P., Mikus, P., Andersson, H. & Ny, T. 1991 The efficiency of the uncleaved secretion signal in the plasminogen activator inhibitor type 2 protein can be enhanced by point mutations that increase its hydrophobicity. *J. biol. Chem.* **266**, 15 240–15 243.

Wagner, L.S., Geddes, J.W., Cotman, C.W., Lau, A.L., Gurwitz, D., Isackson, P.J. & Cunningham, D.D. 1989 Protease nexin-1, an antithrombin with neurite activity, is reduced in Alzeheimer's disease. *Proc. natn. Acad. Sci. U.S.A.* **86**, 8284–8288.

Whinna, H.C., Blinder, M.A., Szewczyk, M., Tollefsen, D.M. & Church, F.C. 1991 Role of lysine 173 in heparin binding to heparin cofactor II. *J. biol. Chem.* **266**, 8129–8135.

Wilson, A.C. & Cann, R.L. 1992 The recent African genesis of humans. *Scient. Am.* April 22–27.

Wright, H.T., Qian, H.X. & Huber, R. 1990 Crystal structure of plakalbumin, a proteolytically nicked form of ovalbumin. *J. molec. Biol.* **213**, 513–528.

Ye, R.D., Ahern, S.M., Le Beau, M.M., Lebo, R.V. & Sadler, J.E. 1989 Structure of the gene for human plasminogen activator inhibitor-2. *J. biol. Chem.* **264**, 5495–5502.

Ye, R.D., Wun, T.-Z. & Sadler, J.E. 1988 Mammalian protein secretion without signal peptide removal. *J. biol. Chem.* **263**, 4869–4875.

Received 3 August 1992; revised 19 February 1993; accepted 2 March 1993

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES