Chloroplastic aspartate aminotransferase from *Arabidopsis thaliana*: an examination of the relationship between the structure of the gene and the spatial structure of the protein

Susan E. WILKIE*, Richard LAMBERT† and Martin J. WARREN*‡

*Department of Molecular Genetics, Institute of Ophthalmology, University College London, 11-43 Bath Street, London EC1V 9EL, U.K., and †Department of Crystallography, Birkbeck College, Malet Street, London WC1E 9EL, U.K.

A clone encoding a plastid isoenzyme of aspartate aminotransferase (AAT5) was isolated from an *Arabidopsis* genomic library and its complete sequence determined. The gene for AAT5 (*asp5*) contains an open reading frame of 2447 bp comprising 11 exons separated by introns ranging in length from 74 to 207 bp. The upstream regulatory region contains a putative TATA box and multiple copies of two sequence motifs, CTCTT and AAAGAT, previously associated with nodule-specific gene activity in legumes. The deduced primary amino acid sequence of the protein product of *asp5* was used to generate a three-dimensional structure of the AAT5 protein by using the computer program Sybyl: Biopolymer Composer and known AAT structures on the protein databases. Both the mature protein and its precursor protein containing a putative N-terminal transit peptide were modelled. The resulting structure of the precursor protein indicated that the transit peptide might also inhibit

dimerization of the protein until after its translocation across the chloroplast membrane. The derived structure of the mature protein was then analysed in terms of its component elements of secondary structure, and the positions on the polypeptide backbone corresponding to intron insertion sites were determined. It is observed that the introns tend to map to regions between structural subdomains of the protein and also map to sites on the surface of the molecule. The *asp5* gene in *Arabidopsis* is thus consistent with Gilbert's exon-shuffling theory of gene evolution [Gilbert (1985) *Science* **228**, 823–824]. A high degree of conservation of intron insertion sites between AAT genes from different plants and animals is observed, particularly within the part of the gene encoding a large β -sheet structure that forms the structural and functional core of the protein. This β -sheet structure is thus believed to comprise an ancient and very highly conserved moiety of the molecule.

INTRODUCTION

Aspartate aminotransferase (AAT; EC 2.6.1.1) catalyses the reversible transamination reaction between aspartate and α -ketoglutarate to give glutamate and oxaloacetate, with pyridoxal 5'-phosphate (PLP) as an essential cofactor. The enzyme plays a key role in the regulation of carbon and nitrogen flux in all organisms and, in eukaryotes, AAT along with malate dehydrogenase comprise a system for transporting reducing equivalents across organellar membranes, the malate–aspartate shuttle.

AAT is the best studied of the PLP-requiring enzymes. It exists as a dimer of two identical subunits of M_r approx. 45000 [1]. In each subunit the PLP cofactor forms a Schiff base with an active site lysine residue (Lys-258 in *Escherichia coli*). The three-dimensional structures of the enzymes for *E. coli* [2–4], chick [5–8] and pig [9] have been determined by X-ray crystallography. This information, together with earlier extensive chemical, kinetic and spectroscopic studies (reviewed in [1]), has resulted in the identification of a number of key residues present in the active site that play a part in the reaction mechanism. More recently, site-directed mutagenesis has been used to examine further the role of some of these key residues, which are found to be highly conserved across all species studied [10–16].

The situation with regard to plant AAT enzymes, however, is not as well advanced as with their bacterial and vertebrate

counterparts. In plants, up to five different isoenzymic forms have been found [17], each being associated with a particular subcellular compartment (e.g. cytosol, chloroplast, mitochondrion). As a result of this multiplicity of isoforms, attempts at purification of the enzymes from plant tissues have not yet allowed crystallization of the proteins and studies of their spatial structures. However, a number of plant cDNA sequences encoding AAT isoenzymes have recently been reported, including five different AAT cDNAs from *Arabidopsis thaliana* [18,19], which may represent the full AAT gene family. One of these *Arabidopsis* sequences, originally referred to as the *Ataat1* gene [19] but here called the *asp5* gene in conformity with the nomenclature used by Schultz and Coruzzi [18], is believed to encode a chloroplastic form of the enzyme. This deduction was based on (i) the close sequence similarity with the alfalfa plastid AAT2 gene [20,21], (ii) the presence of a chloroplast targeting pre-sequence and (iii) the successful importation of the protein product of the gene, the precursor protein, into pea chloroplasts where it was processed into mature AAT.

In addition complete genomic sequences have been determined for the cytosolic and plastid isoenzymes from alfalfa [22], the cytosolic and mitochondrial isoenzymes from proso millet [23] and cytosolic and plastid isoenzymes from lupin [24,25].

In this paper the complete genomic sequence for the *asp5* gene in *Arabidopsis* is reported. In addition a computer-generated

Abbreviations used: AAT, aspartate aminotransferase; PLP, pyridoxal 5'-phosphate.

‡ To whom correspondence should be addressed.

The nucleotide sequence reported will appear in DDBJ, EMBL and GenBank Nucleotide Sequence Databases under the accession number X91865.

model of the spatial structure of the mature AAT protein, based on an analysis of its primary amino acid sequence, is presented. An examination of the internal structure of the gene and of the spatial structure of the protein has allowed certain tentative conclusions to be drawn about the evolution of the gene for AAT.

MATERIALS AND METHODS

Isolation and sequencing of the gene sequence

Whole genomic DNA was prepared from fresh leaf tissue from *A. thaliana* var. Landsberg erecta by the method of Dellaporta et al. [26]. This was used as template to amplify a partial gene sequence of approx. 2 kb by using two gene-specific primers devised from the known cDNA sequence of *Arabidopsis* chloroplastic AAT (WIL1, 5'-GGAAGCTCAGCTTACGT-3'; WIL2, 5'-AGCAGAGTGACTAC-3') [19]. This partial gene sequence was used as a probe to screen a whole genomic library prepared from leaf tissue from *A. thaliana* var. Landsberg erecta (obtained from the *Arabidopsis* DNA Resource Centre, Cologne, Germany) and resulted in the isolation of a positive clone with an approx. 18 kb insert. Southern analysis of DNA isolated from this clone with the same probe as above revealed a single hybridizing band of approx. 5 kb by using the restriction enzymes *Eco*RI and *Kpn*I. This *Eco*RI–*Kpn*I fragment was cloned into pUC18 (pgAt1) and smaller restriction fragments were subcloned from it (pgAt2, pgAt3 and pgAt4 containing a 2.44 kb *Eco*RI–*Hind*III fragment, a 1.03 kb *Hind*III–*Hind*III fragment and a 1.38 kb *Hind*III–*Kpn*I fragment respectively). The complete sequence of the insert in pgAt1 was determined in both senses of the DNA by using T7 DNA polymerase. Sequence analysis was performed with the University of Wisconsin gcg software package (CLRC Daresbury Laboratory, Daresbury, Warrington, Cheshire, U.K.).

Computer modelling of the spatial structure of the *Arabidopsis* AAT protein

Modelling of the structure of the mature *Arabidopsis* chloroplastic AAT exploited the high sequence similarity (up to 49% identity) between the deduced *Arabidopsis* primary sequence and the primary sequences of known AAT structures available in the Brookhaven protein structure database [27].

Initially the Brookhaven protein structure database was interrogated for three-dimensional AAT structures. The structures extracted were analysed with the program Procheck (CCP4 package CLRC Daresbury Laboratory). Clearly the quality of the stereochemical geometry of the model would closely reflect the quality of the input structures. Homology modelling was conveniently implemented in the Composer option [28,29] of the Sybyl software suite (Tripos Associates). By using a database created for Sybyl from AAT structures extracted from the Brookhaven database and the deduced primary sequence of the *Arabidopsis* chloroplastic AAT, the program performed multiple sequence alignments [30] with a default gap penalty of 8. The best pair of aligned sequences were selected and used to define regions of high sequence similarity. The program identified groups of topologically equivalent residues, which were used to generate an optimal structural alignment for the sequences, and a framework for the model was created based on the structurally conserved regions. Having organized the sequence into structurally conserved regions and variable regions and having built the framework, the program examined the geometric constraints required to insert loops between the structurally conserved regions. Possible candidate loops for insertion were recruited from the

database of AAT structures. The loops or fragments acquiring the highest score with selection criteria based on geometric and sequence discrimination were appended to the framework of the model. Variable regions not satisfied by loops from the AAT database were then fitted with loops from a larger, more comprehensive, Sybyl database of three-dimensional structures (pdatables).

Energy minimization routines (with Tripos force fields; convergence criterion 0.05 kcal/mol per Å) within the Sybyl software allowed the strain between the structurally conserved regions of the framework and the inserted variable loops to be relieved. The final minimized model was checked for stereochemical conformity by using Procheck (CCP4 Package). Further tests consistent with a correctly folded protein were performed: maintenance of hydrophobic cores, exposure of inappropriate hydrophobic residues to solvent, and the presence of charged or polar residues in the interior of the protein causing disruption of the hydrophobic cores of the domains. These tests were also performed after cofactor insertion and dimer assembly [31].

The molecular graphics program O was used for the assembly of the dimer, insertion of the cofactor and superimposition of the model on to other AAT structures [32]. Interatomic distances between residues in the active site and the cofactor were estimated with Contact (CCP4 Package). The spatial structure of a single subunit was analysed in terms of its component elements of secondary structure by using DSSP [33].

RESULTS AND DISCUSSION

The *asp5* gene from *Arabidopsis*

A partial *asp5* gene sequence was amplified from *Arabidopsis* whole genomic DNA with PCR primers designed from the cDNA sequence [19]. This was used to probe an *Arabidopsis* whole genomic DNA library and resulted in the isolation of a positive clone, pgAt1, containing a 4836-base insert. The AAT gene sequence contained within clone pgAt1 comprises an open reading frame of 2447 base pairs, 2236 base pairs of upstream sequence and 153 base pairs of downstream sequence (see Figure 1). Thus although the sequence is likely to contain most of the promoter elements and the coding sequence, the transcriptional stop sequence and poly(A) addition signal may be missing.

In the regulatory region upstream from the ATG start codon, a putative TATA box (TATAAA) is located at position 1821. In addition this region is found to contain multiple copies of two sequence motifs, CTCTT (five copies, one in reverse orientation) and AAAGAT (four copies, three in reverse orientation), which are conserved in several legume nodule-enhanced genes including the genes encoding alfalfa AAT2 [22], the genes for AAT1 and AAT2 from lupin [24,25] and the genes for leghaemoglobin and nodulin from soybean [34–36]. These sequence motifs have been shown to exert a positive effect on nodule-specific expression in soybean and have been proposed to represent nodule-specific *cis*-acting control elements. Their presence in the gene promoter from a non-nodulating species such as *Arabidopsis* might, however, indicate that they have a more general gene-enhancing function and are not nodule-exclusive. On the other hand the sequences are very short (five or six bases) and their presence could be of doubtful statistical significance. Moreover the relative positions of such motifs within a promoter might be important for regulatory activity. Thus it would be unwise to speculate further on their functional significance in this *Arabidopsis* promoter.

The protein-encoding part of the gene consists of 11 exons, separated by 10 introns ranging in length from 74 to 207 base

GGTACCAAAACAAACACAAGTACTAATCATTTTACTTTTACCGGATCTCTGATATCTGATTTTATCATCTAAATCTCAGAAATGAAGTTGATTGTTGAGCAAGATTCACCACCCGA	120
ACATCATCTCATTTGTTTGGATATGGAATGAATCAGTTCGAGTTTATCGTCTACGAGCTGATGAAAGCGGATCATTTGGATACACAGTTACACGGTAAAAATAGCTCAAAACATCTTC	240
TCATCATCGTGTAAATAGATCATGAGTTAAATTAATACTATAGTATAACAGAGATAAAACAAATCGATGTGCCACATGAAATAAATAATTTTGTITTTGATAAAATTTAATGTTTGATC	360
ATTAAATGGTTAAGGACCTTCTCGGGGATCGGCTTTAACAATGGCACATGCGGATGAAGATTGCTCTTGATACAGCAAGGTACTACTCAAGTGGCCACATTTAACAATCCTCTCTACCGG	480
TTGATTAATTTTCGGTTTAGGTAAACCGTTTCTTATGTGTGTTTGGAAATAGAGCTGTGTAGTATCTCCACGAGCGTTGTCGTCCTCCGGTTATCCACAGAGATCTTTAAATCGTCAATA	600
TTCTCCTTGATTTCTCTTCAACGCCAAGGTAAACAACCTTGAATCAITTAATGATGATTCTGTCTGAGCTAACCGAACAAGATTGTTCGTGTGATTCAGAATCATTTCAITTCATTTGT	720
CTTTAGTTCCATCTCTAACGATAGATCAACATTCAGATTTCGGATTTTGGTCTTTCGGGTAAATGGTGGGGCTCAGCGCAAAACACATTTAACTATCAGGAACACTTGGTTATGTTGCT	840
CCAGAATATCTCTAGATGGTAAATATTAACATTTCCCTCTATATGATCGACTTTGACTCGTTGTCTATATCACTAGAGTTATAGCACATAGTTTGATAGTCATAATCATGTGACCCCA	960
TAACATTAAGTTTCTTTTCATTACGTCTAAATTTGTGTCTAAATCAAAATCAATTTTGTGTGTGATGATTAATTTCAACTAATCACACATAAAATCATTAATTTGATTAGAGATGGTAGT	1080
TGTTGGGTCGTAAACCGTTTATTTAGTTTGTCTCATTTGCAACATCTTTATGTCCTCAGGAAATTTGACGGATAAGAGTGATGTTTATGCGTTTGGTGTGGTTTACTTTGAATCTTTG	1200
TTAGGAAGACGGCCGGTTGAGAAATTTAGTTTCGGTTTCAGTGTCAATCTCTTTGTCACCTTGGGTAAAGTCCGGTTTGTGTGACGTTTATGCGCATCAATACATGAATCAGATAGCATTTGTTA	1320
ATGAGTGTACCGTTTCGAAAAATTTGAGCAATGCCCCAATTCAGGATAGATCAAAAGCTTCGAAAAATCGTGGATCCGGTTTATCAAGATACAAATGATCAATGACACTTATACAGGT	1440
TTGCGTCACTCTTTTGTCTTTTGTGTTTAGGCTTAAAAATCCGATTTGGTTACATATTTTGTGTGTGAAACGTAATAAAACCGGAAATAAATTTGTTGAGGTGGCAGCGTGGCAGTGCTTT	1560
GTGTACAAACGAAACCGAGTTATCGACCGTTGATAACCGATGTTCTTCACTCACTAGTTCCATTGGTTAAGGTAGAGCTAGGAGGACTCTCCGGTTAATACCATCATCGTCTTGATTTA	1680
AGAAATAATATTTTGTGCTTCTTTTGAATTTTACAGTTTGATCTCTGATGCTTTGAGTAATTTAATTTGGGGGTGAAAAATGTGAGGACAGAAGGGTATTCACTTGTCAITTAATG	1800
TATTTGTAAATTTGATTAATTTATAAATTTGTGTGTGCTAAGTGTAAATATTCATTAATTTGTCATGAACCACTAATCACAATCCACTAGTATCCATCCATCCACTCTTTCAACTTTT	1920
AAACATAATCTAAATGACTATCTTACCCCTATCAAGACAAGCTGAGAAGTGAGAACACTTCGTAGACATCGTCGTCTCTCTCTCGTTGAATCTTGTGTTAATCTTTTCTTCAATCT	2040
CTCGATCTGATCGTTATCGAAACCGCGGATTTCTAACCAAGTCCAGtgacaaatcagctctgctaaatactcacaaagtctcgtgatcatatcttggatcacatcattgtcttatt	2160
tttggaaattgcttatctcgcactagttcactgatttgagatctgtgcagTTAATTTGTGTCTCATAGCATTCCATGGCTTCTTTAATGTTATCTCTCGTTCCACTTCTCTGTTACC	2280
GCGCGAGATTAAACAAGgtacttttcgcgttctatcgagtatatggttagaattacgcacatgatttttcaaaatcttactttttgttttcgattaaacttgaacagGATTAACGTAAAG	2400
oArgGluIleAsnLys CTTTGGAATCTCTGCTTTCGAACCGTTCTCTAAAGCAAGgtactctcttcttggtagtgcgaatctgaaccatatgaacaattttgaatgacgttttagagtgttattagttctaa	2520
LeuGlyThrSerAlaSerAsnProPheLeuLysAlaLys taagtgtagtttgcattcgaattttggtgattttggtttcagTCTTTTAGCAGAGTGACTATGACGGTTGCAGTGAAGCCTTCTCGTTTCGAGGGGTATAACTATGGCTCCACCAGACCT	2640
SerPheSerArgValThrMetThrValAlaValLysProSerArgPheGluGlyIleThrMetAlaProProAspPro ATTCTTGGAGTCAGTGAAGCATTTCAAGCTGACATAACGGGATGAACTCAATCTTGGTGTGGTGTCTTATCGTACTGAGGAATCCAGCCTTATGCTTAAATGTTGTTAAAAAGgtt	2760
IleLeuGlyValSerGluAlaPheLysAlaAspThrAsnGlyMetLysLeuAsnLeuGlyValGlyAlaTyrArgThrGluGluLeuGlnProTyrValLeuAsnValValLysLys ggaactgttaactctagtttaactcgtatggtcctttcattgtagttaatttcattactgtcgttcttagCGGAGAAATTTGATGTTGAGAGAGGAGATAACAAAGAGgttaacttggt	2880
AlaGluAsnLeuMetLeuGluArgGlyAspAsnLysGlu ttatttgaattgtgcaagcgattacatcatcagaatatcactaaattatacctcacatggatggaacttgcagTATCTTCCAAATGAGGGGTGGCAGCATTCACAAGGCTACTGCTG	3000
TyrLeuProIleGluGlyLeuAlaPheAsnLysAlaThrAlaG AGTTGCTATTTGGAGCTGGTCACTCTGTTATTAAGGAACAAAGagtaattctgcatcttttggatcatgttattaaaaatgttttctcatgattgttctgtgagtaataccattt	3120
luLeuLeuPheGlyAlaGlyHisProValIleLysGluGlnArg accttggtctctaaatcaccagGTAGCAACAATTCAGGCTCTTTTCGGGAACAGGTTCACTGCGATTAGCAGCGGCTCTTATAGAGCGTTATTTCCCTGGAGCAAAAGTTGTGATCTCATCA	3240
ValAlaThrIleGlnGlyLeuSerGlyThrGlySerLeuArgLeuAlaAlaLeuIleGluArgTyrPheProGlyAlaLysValValIleSerSer CCAACCTGGGgtacattgtcttgaccacaacatttttgagtatttgggttcatatttaaccgctttatcaactgacaaaataccacatcagGTAATCACAAGATATCTTCAATGAT	3360
ProThrTrpG GCCAAAGTTCCGTGGTCCGAATACCGCTACTATGATCCAAAAACAATTTGGTTTGGATTTTGAGGGAATGATAGCAGATATAAAGgtttgttgcctagaaggctatgttttagttgaca	3480
AlaLysValProTrpSerGluTyrArgTyrTyrAspProLysThrIleGlyLeuAspPheGluGlyMetIleAlaAspIleLys ttgcttttatatgatcatcaacaacaggatatttattgtgatgcaaaagttagtatttggtttcagttatactgtctgtgtgataaaacgagtaacttgtttaaagacaactgaaacc	3600
ctgatggattataatggttaattcttgaatttctgttatgtaaaatgtacagGAAGCTCCAGAAGGATCCTTCATCTTGTCTTACCGATGTCTCACAACCAACTGGAATTGACCCAACA	3720
GluAlaProGluGlySerPheIleLeuLeuHisGlyCysAlaHisAsnProThrGlyIleAspProThr CCAGAACAGTGGGTAAAAATTTGCTGATGTCTTCAGGAAAGAACCATATCCCATTTTTCGATGTGTGATACCAAGgtatccccctattcctacatttccgaatctcgtgtttccgagaga	3840
ProGluGlnTrpValLysIleAlaAspValIleGlnGluLysAsnHisIleProPhePheAspValAlaTyrGln agtaagcgaaacatattatgtaacgctacatgcacctattttccagGGCTTTGCTAGTGGGAAGCCTTGATGAAGATGCAGCATCTGTGAGATTATTTGCTGAGCGGGAAATGGAGTTT	3960
GlyPheAlaSerGlySerLeuAspGluAspAlaAlaSerValArgLeuPheAlaGluArgGlyMetGluPheP TTGTTGCTCAGTCAATAGTAAAAATTTAGGTTTGTATGCGAAAGAAATTTGGGGCAATCAATGTCTGTCTCTCAGCTGATGCTGCTACAAGTaaacttactctaaattttcttat	4080
heValAlaGlnSerTyrSerLysAsnLeuGlyLeuTyrAlaGluArgIleGlyAlaIleAsnValValCysSerSerAlaAspAlaAlaThrAr catgagcctgttgcaagtgtatcctggagctcagtttctatctttaccgtcacgaaccatatctagctacaggttttatggaacctttaccagtaaacactacttgccttcagGGTCAAGA	4200
gValLysS GCCAGTTGAAAAGGATTGCTCGGCTATGTACTCGAATCCACCAGTTTCATGGGGCAGAAATCGTGGCCAAATGTAGTGGGTGATGTAACATATGTTTCAAGTGAATGGAAAGCAGAGATGGA	4320
erGlnLeuLysArgIleAlaArgProMetTyrSerAsnProProValHisGlyAlaArgIleValAlaAsnValValGlyAspValThrMetPheSerGluTrpLysAlaGluMetGluM TGATGGCAGGAAGATAAAGACGGTTAGACAAGAGCTGTATGATAGCCTCGTTTCAAAAGACAAGAGCGGGAAGGACTGGTCTCTCATTTCTGAAGCAAAATTTGGCATGTTCTTTACCG	4440
etMetAlaGlyArgIleLysThrValArgGlnGluLeuTyrAspSerLeuValSerLysAspLysSerGlyLysAspTrpSerPheIleLeuLysGlnIleGlyMetPheSerPheThrG GCCTAAACAAAGCTCAGgtatgccaacaattttaatactacacagattccctttggccttggtagtctgtgttgaagtgtgtgttcttcttggtaacgagcagAGCGATAAC	4560
lyLeuAsnLysAlaGln ATGACGGCAAAATGGCATGTGTATATGACTAAAGACGGGAGATATCATTTGGCCGGATTATCTCTGGCCAAATGCGAGTATCTTGTGATGCGATCATCGACTCATACCATTAACGTAAAGC	4680
MetThrAspLysTrpHisValTyrMetThrLysAspGlyArgIleSerLeuAlaGlyLeuSerLeuAlaLysCysGluTyrLeuAlaAspAlaIleIleAspSerTyrHisAsnValSer TGAGCTTCCATCTCAGTAGATGACAATAAGAACAGTTTATATACCTTTTATAGCTTCTTCTGATCTTTGATTGAACACAGATGGCTAAATTAATAGCTAGCATCACACTATGATTG	4800
... CTAGACAAATTTCAAGGAACTTTTGTATTGAAATTC	4836

Figure 1 Nucleotide sequence of the *asp5* gene in *Arabidopsis*

The translational start (ATG) and stop (TGA) codons are shown in bold. The introns are shown in lower case. The deduced amino acid sequence of the precursor protein encoded by the exons is given below the nucleotide sequence. In the upstream regulatory region, the putative TATA box (TATAAA) is boxed and copies of possible regulatory motifs CTCTT and AAAGAT (and their complements) are underlined.

Figure 2 Representations of the computer-generated model of *Arabidopsis* AAT5

(A, B) Two representations of a single subunit of AAT5 taken from different aspects. In (A) α -helices are represented as ribbons (pink), β -strands as wide arrows (yellow) and loops as thin rope. In (B) the polypeptide backbone is represented as a simple ribbon that has been coloured to show segments encoded by different exons: exon III (N-terminus), blue; exon IV, cyan; exon V, green; exon VI, purple; exon VII, orange; exon VIII, red; exon IX, white; exon X, blue/green; and exon XI (C-terminus), yellow. (C) A line representation of the polypeptide backbone of the *Arabidopsis* homodimer (in blue) superimposed on the known structure of the open conformation of chicken mitochondrial AAT (in red). Where the superimposition is perfect a purple colour is generated. The cofactor (PLP) is shown in green. (D) A superimposition of the *Arabidopsis* precursor protein (in blue) on the mature dimer (in red). That part of the precursor protein corresponding to the mature protein superimposes perfectly to produce a purple colour whereas the transit sequence is clearly revealed in blue.

pairs. A further intron of 124 base pairs interrupts the transcribed but not translated sequence upstream from the translational start codon. All the introns conform to the splice site consensus

GT ... AG [37]. The sequence of the coding region of the gene is identical with the reported cDNA sequence [19].

Because the gene encodes the chloroplastic isoenzyme of AAT,

Figure 3 Topology diagram showing the secondary and tertiary structural elements in the model of a single subunit of *Arabidopsis* AAT

α -Helices are shown as cylinders and β -strands as arrows. The residues marking the ends of these elements are numbered. The cofactor is shown covalently linked to Lys-298.

the protein deduced from it would be expected to incorporate a transit sequence at its N-terminus. A sequence comparison with the cytosolic AAT enzyme from *Arabidopsis* [18] leads to the deduction that this presumed transit sequence consists of approximately the first 50 amino acid residues. This presumed transit peptide is therefore encoded in the first two exons and a portion of the third exon (see Figure 4).

Computer-generated model of the spatial structure of the *Arabidopsis* AAT protein

Because the identity between the primary sequences of the plant and bacterial AATs is approx. 50 %, it was thought that a 'plant' AAT structure could be modelled from the pre-existing structures for AAT on the databases. Thus the deduced primary sequence of mature *Arabidopsis* chloroplastic AAT was used to generate a three-dimensional structure of the protein with the computer software package Sybyl: Biopolymer Composer (Tripos Associates) and information in the Brookhaven protein structure database. With the known structure of chicken mitochondrial AAT as a blueprint [7] the computer generated the structure represented in Figures 2(a) and 2(b) for a single AAT subunit.

The overall structure of the *Arabidopsis* AAT model is similar to the known structures of the *E. coli* and animal AATs. Each subunit contains 15 α -helices and 11 β -strands. These fold into a structure comprising a large and a small domain and an extended N-terminal arm, the end of which interacts with the other subunit. The small domain is formed by two parts of the polypeptide chain, the first comprising residues from the N-terminus to Pro-94 and the second from Arg-369 to the C-terminus. The large domain is formed by the remaining residues. The two domains are connected by a long α -helix consisting of 31 residues (Val-353 to Ser-383) with a kink at Arg-369. This kink is also found in the structures of mammalian AATs and is

believed to serve as a hinge for the movement of the small domain in the conformational change that occurs on substrate binding [38].

A wide β -sheet structure comprising seven of the β -strand elements [six parallel (p) and one anti-parallel (a)] forms the core of the large domain. The arrangement of these β -strands is pppppap. All seven of the β -strands have a tendency to twist right-handedly to some extent and the pleated β -sheet formed by them assumes a left-handed twist when viewed along the β -sheet normal to the strands. As in the mammalian and *E. coli* AATs, this β -sheet structure is constructed from a series of simple repeated structural motifs consisting of α -helix- β -strand, except that no α -helix is apparent between the sixth and seventh β -strands. However, the polypeptide backbone executes a sharp twisting turn in this region reminiscent of the small (five-residue) α -loops found in the other AAT structures [38]. This turn incorporates Lys-298, which binds the pyridoxal phosphate cofactor. The β -sheet structure as a whole exhibits similarity with the dinucleotide binding domain of dehydrogenases [39,40]. In the dehydrogenases two $\beta\alpha\beta$ -folds form hydrophobic pockets that accommodate the adenosine and nicotinamide aromatic residues of the dinucleotide cofactor [41]. The β -turn- β motif in the *Arabidopsis* AAT structure, which accommodates the aromatic nucleus of the PLP cofactor, shows similarity to these $\beta\alpha\beta$ -folds.

The small domain comprises four short β -strands and five α -helices. The β -strands interact via a few hydrogen bonds to form two small regions of β -sheet.

The structure of chicken cytosolic AAT was found to contain five hydrophobic clusters, three in the large domain and two in the small domain, which contribute significantly to the overall stability of the structure [8]. Our *Arabidopsis* AAT structure contains hydrophobic regions that can be subdivided into five similar clusters.

Figure 4 Correlation of exon structure of the *asp5* gene in *Arabidopsis* with the structural and functional domains of the protein

The positions of introns in the *asp5* gene are mapped on to the amino acid sequence of the enzyme. The amino acid residues that correspond to or abut introns are indicated. The locations and sizes of secondary structural elements within the protein are shown as open boxes (α -helices 1–15) or black boxes (β -strands 1–11). Intron positions marked with an asterisk are highly conserved in other eukaryotic AAT genes.

The AAT enzyme itself exists as a homodimer, and Figure 2(C) illustrates how two of the above subunits might fit together. In Figure 2(C) the modelled structure of the *Arabidopsis* enzyme has been superimposed on the known structure of the chicken mitochondrial enzyme (the open conformation) to allow a gross comparison of the two structures to be made and to enable regions of dissimilarity to be examined. A high degree of overlap of the two structures is observed and any minor differences in the disposition of the polypeptide backbone of the two proteins occur on the surface of the molecule and are distant from the active site.

The dimer contains two equivalent active sites, each formed from the residues of both subunits. As a test of the reliability of the modelled structure, the pyridoxal phosphate cofactor was positioned within the predicted active site region bound to Lys-298 via a covalent imino linkage between C-4' of the cofactor and the ϵ -amino group of the lysine residue. The relative positions of other residues in the active site known to interact with the cofactor were then determined. It was observed that the carboxylate group of Asp-262 is within hydrogen-bonding distance of the pyridinium NH of the cofactor and also of the imidazole NH of His-184. Residues Trp-181, His-184, Asp-262 and Tyr-265 together form a pocket for the pyridoxal C-2' methyl and C-3' hydroxyl groups, with the 3' oxygen of the latter within hydrogen-bonding distance of the phenolic hydroxyl of Tyr-265. The indole ring of Trp-181 is observed to stack parallel to the pyridinium ring of the cofactor. Ser-153, Thr-155, Ser-295 and Arg-306, together with Tyr-116 from the second subunit, together form a pocket for the 5'-phosphate of the cofactor capable of stabilizing its double negative charge via hydrogen bonds and other bonding interactions. All these interactions are believed to be important for stabilization of the cofactor and to orientate it correctly within the active site. The observation that these interactions are satisfied by the model gives credence to its

reliability. However, until crystals of the *Arabidopsis* enzyme have been obtained and its actual structure determined, it is not possible to say how accurately the computer-generated structure models the actual structure.

Having generated a model for the mature protein, Composer was then used to model the structure of the AAT precursor protein. One of the benefits of this model was that it allowed attention to be focused on the location and conformation of the putative transit peptide, permitting speculation about its functional significance. Figure 2(D) shows a superimposition of the model of the precursor protein (in blue) on the structure of the homodimer (in red). Where the two structures overlap a purple colour is generated. This serves to highlight the putative transit peptide (which remains in blue). It is observed that the transit sequence seems to penetrate deep into the space occupied by the second subunit of the dimer. This would suggest that the presence of the transit sequence will inhibit dimerization of the precursor protein. The primary function of the transit sequence is to target the protein to the chloroplast and to facilitate its transfer across the chloroplast membrane. Our observation implies that a second function of the sequence may be actively to inhibit dimerization and subsequent generation of an active enzyme until after translocation into the chloroplast.

Analysis of the spatial structure of the *Arabidopsis* AAT model and the structure of the *asp5* gene

The spatial structure of the *Arabidopsis* AAT model was analysed in terms of its component elements of secondary structure and the results were formulated into a topology diagram presented as Figure 3. By using a ribbon representation of the structure, sections of the polypeptide backbone corresponding to different exons in the gene were colour-contrasted as shown in Figure 2(B) so as to allow a visual analysis of the relationship between the

elements of secondary structure and the exon-exon splice junctions in the gene. The result of this analysis is given in Figure 4.

It is observed that five of the eight splice junctions in the part of the gene encoding the mature protein map to polypeptide loops between elements of secondary structure rather than within the secondary structural elements themselves. No splice junctions map within any of the β -strand elements. In particular, within the segment of the polypeptide forming the β -sheet structure of the large domain, none of the secondary structural elements is interrupted by splice junctions. This observation is consistent with Gilbert's exon theory of gene evolution [42], whereby primordial exons encoded functional protein subdomains and the ancient proteins ancestral to our modern proteins evolved from them by a process of exon shuffling. It would seem that within the *asp5* gene in *Arabidopsis* the relics of Gilbert's primordial exons can still be discerned, especially within the part encoding the large β -sheet structure of the large domain. It could be argued that the β -sheet structure represents the structural and functional core of the molecule, incorporating the cofactor-binding site (Lys-298) and most of the key residues conferring catalytic activity. It would therefore be expected to be the most rigidly conserved region. The conservation of exon-encoded subdomains in this region would therefore be more likely than in other less functionally rigid regions of the molecule. Branden [43] proposed that genes for proteins with β -strand- α -helix structures were constructed by fusing together DNA sequences encoding single β -strand- α -helix motifs. The results presented here do not seem to support this proposition, however, as the splice junctions do not divide this region of the protein into multiples of such motifs.

It is also observed that in the *Arabidopsis* AAT model the exon splice junctions tend to map to the surface of the protein (see Figure 2B). Possible exceptions are intron 6 (Gly-182), which seems to map within the pocket between the N-terminal arm and the large domain, and intron 8 (Gln-266-Gly-267), which maps to within a surface groove in the large domain. However, even these positions might conceivably be within contact with the solvent. The localization of splice junctions on the protein surface has also been observed in the mouse cytosolic and mitochondrial genes [44]. Craik et al. [45] argued that during the evolutionary development of families of related genes (such as the AAT isoenzyme family), the presence of introns provided an enhanced possibility for variation of the gene product if such introns mapped to the protein surface rather than to the central core region, where changes would be more likely to exert a negative effect (i.e. loss of function).

A phylogenetic analysis based on the protein sequences of a number of plant and animal AAT isoenzymes revealed that corresponding isoenzymes from different cellular locations tend to group together, and five major subgroups can be discerned: animal cytosolic, animal mitochondrial, plant cytosolic, plant mitochondrial and plant plastid [22,23]. An unusually high degree of conservation of the positions of exon splice junctions was observed in genes encoding AAT isoenzymes, both within and between these subgroups, and it is believed that this indicates that genes encoding all AAT isoenzymes were derived from a common ancestral gene that contained introns. A comparison of the splice junctions in the *asp5* gene from *Arabidopsis* with those in AAT genes from the other subgroups reveals that five of the eight splice junctions present in the part of the gene encoding the mature protein correspond exactly with splice junctions in other AAT genes, and a sixth represents a slippage of only two codons (Figure 4). Within the region of the gene encoding the β -sheet structure (from Arg-145 to Cys-314) this conservation of splice junctions is particularly strong and exact counterparts for each

of introns 5 to 8 may be found. This observation further supports the proposition that the β -sheet structure represents an ancient and very highly conserved moiety of the protein.

We thank the Department of Crystallography, Birkbeck College, London, for their generous provision of computing facilities, and Dr. Stephen P. Woods for his reading of the manuscript. Financial support by a grant from the Biotechnology and Biochemical Sciences Research Council, U.K., is gratefully acknowledged.

REFERENCES

- Christen, P. and Metzler, D. E. (eds.) (1985) *Transaminases*, John Wiley and Sons, New York
- Smith, D. L., Almo, S. C., Toney, M. D. and Ringe, D. (1989) *Biochemistry* **28**, 8161-8167
- Jager, J., Moser, M., Sander, U. and Jansonius, J. N. (1994) *J. Mol. Biol.* **239**, 285-305
- Okamoto, A., Higuchi, T., Hirotsu, K., Kuramitsu, S. and Kagamiyama, H. (1994) *J. Biochem. (Tokyo)* **116**, 95-107
- Borisov, V. V., Borisova, S. N., Kachalova, G. S., Sosfenov, N. I. and Vainshtein, B. K. (1980) *Nature (London)* **284**, 189-190
- Ford, G. C., Eichele, G. and Jansonius, J. N. (1980) *Proc. Natl. Acad. Sci. U.S.A.* **77**, 2559-2563
- McPhalen, C. A., Vincent, M. G. and Jansonius, J. N. (1992) *J. Mol. Biol.* **225**, 495-517
- Malashkevich, V. N., Strokopytov, B. V., Borisov, V. V., Dauter, Z., Wilson, K. S. and Torchinsky, Y. M. (1995) *J. Mol. Biol.* **247**, 111-124
- Arnone, A., Rogers, P. H., Hyde, C. C., Briley, P. D., Metzler, C. M. and Metzler, D. E. (1985) in *Transaminases* (Christen, P. and Metzler, D. E., eds.), pp. 138-154, John Wiley and Sons, New York
- Kamitori, S., Okamoto, A., Hirotsu, K., Higuchi, T., Kuramitsu, S., Kagamiyama, H., Matsuura, Y. and Katsube, Y. (1990) *J. Biochem.* **108**, 175-184
- Goldberg, J. M., Swanson, R. V., Goodman, H. S. and Kirsch, J. F. (1991) *Biochemistry* **30**, 305-312
- Inoue, K., Kuramitsu, S., Okamoto, A., Hirotsu, K., Higuchi, T. and Kagamiyama, H. (1991) *Biochemistry* **30**, 7796-7801
- Planas, A. and Kirsch, J. F. (1991) *Biochemistry* **30**, 8268-8276
- Pan, Q.-W., Tanase, S., Fukumoto, Y., Nagashima, F., Rhee, S., Rogers, P. H., Arnone, A. and Morino, Y. (1993) *J. Biol. Chem.* **268**, 24758-24765
- Ziak, M., Jager, J., Malashkevich, V. N., Gehring, H., Jaussi, R., Jansonius, J. N. and Christen, P. (1993) *Eur. J. Biochem.* **211**, 475-484
- Kohler, E., Seville, M., Jager, J., Fotheringham, I., Hunter, M., Edwards, M., Jansonius, J. and Kirschner, K. (1994) *Biochemistry* **33**, 90-97
- Ireland, R. J. and Joy, K. W. (1985) in *Transaminases* (Christen, P. and Metzler, D. E., eds.), pp. 376-384, John Wiley and Sons, New York
- Schultz, C. J. and Coruzzi, G. M. (1995) *Plant J.* **7**, 61-75
- Wilkie, S. E., Roper, J. M., Smith, A. G. and Warren, M. J. (1995) *Plant Mol. Biol.* **27**, 1227-1233
- Gantt, J. S., Larson, R. J., Farnham, M. W., Pathirana, S. M., Miller, S. S. and Vance, C. P. (1991) *Plant Physiol.* **98**, 868-878
- Robinson, D. L., Kahn, M. L. and Vance, C. P. (1994) *Planta* **192**, 202-210
- Gregerson, R. G., Miller, S. S., Petrowski, M., Gantt, J. S. and Vance, C. P. (1994) *Plant Mol. Biol.* **25**, 387-399
- Taniguchi, M., Mori, J. and Sugiyama, T. (1994) *Plant Mol. Biol.* **26**, 723-734
- Kirk, C. A., Mett, V. and Reynolds, P. H. S. (1994) *Plant Physiol.* **105**, 763-764
- Mett, V., Mett, V. L. and Reynolds, P. H. S. (1994) *Plant Physiol.* **106**, 1683-1684
- Dellaporta, S. L., Wood, J. and Hicks, J. B. (1983) *Plant Mol. Biol. Rep.* **14**, 19-21
- Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rogers, J. R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) *J. Mol. Biol.* **112**, 535-542
- Blundell, T. L., Carney, D. P., Gardner, S., Hayes, F. R. F., Howlin, B., Hubbard, T. J. P., Overington, J. P., Singh, D. A. and Sibanda, D. L. (1988) *Eur. J. Biochem.* **172**, 513-520
- Sutcliffe, M. J., Haneef, I., Carney, D. and Blundell, T. L. (1987) *Protein Eng.* **1**, 377-384
- Needleman, S. and Wunsch, C. (1970) *J. Mol. Biol.* **48**, 443-453
- Novotny, J., Rasin, A. A. and Bruccoleri, R. E. (1988) *Proteins* **4**, 19-30
- Jones, T. A. and Thirup, S. (1986) *EMBO J.* **5**, 819-822
- Kabsch, W. and Sander, C. (1983) *Biopolymers* **22**, 2577-2637
- Sandal, N. N., Bojsen, K. and Marcker, K. A. (1987) *Nucl. Acids Res.* **15**, 1507-1519
- Stougaard, J., Sandal, N. N., Gron, A., Kuhle, A. and Marcker, K. A. (1987) *EMBO J.* **6**, 3565-3569
- Jorgensen, J.-E., Stougaard, J. and Marcker, K. A. (1991) *Plant Cell* **3**, 819-827
- Breathnach, R. and Chambon, P. (1981) *Annu. Rev. Biochem.* **50**, 349-383

- 38 Jansonius, J. N., Eichele, G., Ford, G. C., Picot, D., Thaller, C. and Vincent, M. G. (1985) in *Transaminases* (Christen, P. and Metzler, D. E., eds.), pp. 110–138, John Wiley and Sons, New York
- 39 Rao, S. T. and Rossmann, M. G. (1973) *J. Mol. Biol.* **76**, 241–256
- 40 Rossmann, M. G., Moras, D. and Olsen, K. W. (1974) *Nature (London)* **250**, 194–199
- 41 Wierenga, R. K., De Maeyer, M. C. H. and Hol, W. G. J. (1985) *Biochemistry* **24**, 1346–1357
- 42 Gilbert, W. (1985) *Science* **228**, 823–824
- 43 Branden, C.-I. (1986) in *Current Communications in Molecular Biology: Computer Graphics and Molecular Modelling* (Flatterick, R. and Zoller, M., eds.), pp. 45–51, Cold Spring Harbor Laboratory, Cold Spring Harbour, NY
- 44 Tsuzuki, T., Obaru, K., Setoyama, C. and Kazunori, S. (1987) *J. Mol. Biol.* **198**, 21–31
- 45 Craik, C. S., Rutter, W. J. and Fletterick, R. (1983) *Science* **220**, 1125–1129

Received 7 February 1996/24 June 1996; accepted 28 June 1996