Common sequence motifs coding for higher-plant and prokaryotic *O*-acetylserine (thiol)-lyases: bacterial origin of a chloroplast transit peptide?

Norbert ROLLAND, Dominique JOB and Roland DOUCE*

Unité Mixte C.N.R.S./Rhône-Poulenc Agrochimie (U.M. 41) du Centre National de la Recherche Scientifique, Rhône-Poulenc Agrochimie, 14–20 rue Pierre Baizet, 69263 Lyon Cédex 9, France

A comparison of the amino acid sequence of *O*-acetylserine (thiol)-lyase (EC 4.2.99.8) from *Escherichia coli* and the isoforms of this enzyme found in the cytosolic and chloroplastic compartments of spinach (*Spinacia oleracea*) leaf cells allows the essential lysine residue involved in the binding of the pyridoxal 5'-phosphate cofactor to be identified. The results of further sequence comparison of cDNAs coding for these proteins are discussed in the frame of the endosymbiotic theory of chloroplast

evolution. The results are compatible with a mechanism in which the chloroplast enzyme originated from the cytosolic enzyme and both plant genes originated from a common prokaryotic ancestor. The comparison also suggests that the 5'-non-coding sequence of the bacterial gene was transferred to the plant cell nucleus and that it has been used to create the N-terminal portions of both plant enzymes, and possibly the transit peptide of the chloroplast enzyme.

INTRODUCTION

Plants and micro-organisms can perform the *de novo* synthesis of sulphur amino acids, whereas animals require a dietary source of methionine or its precursor, homocysteine. *O*-Acetylserine (thiol)-lyase (EC 4.2.99.8), a pyridoxal 5'-phosphate (PLP)-dependent enzyme, is a component of this metabolic pathway. The enzyme catalyses the final step of L-cysteine biosynthesis (i.e. the synthesis of L-cysteine from *O*-acetyl-L-serine and sulphide). Our understanding of this pathway comes primarily from studies conducted with the bacterial enzymes. Both *Salmonella typhimurium* and *Escherichia coli* contain two isoforms of this enzyme, called A and B [1–3]. Unlike the A isoform, the B enzyme also catalyses the formation of S-sulphocysteine from *O*-acetyl-L-serine and thiosulphate [4–5]. Both isoforms are homodimers of molecular mass 60–68 kDa that contain one molecule of PLP per enzyme subunit. The nucleotide sequences encoding these proteins are known [6–8].

Recent studies proved the existence in higher plants of several isoforms of *O*-acetylserine (thiol)-lyase, each of which is specific for each cell compartment involved in protein biosynthesis, namely, plastids, mitochondria and cytosol [9,10]. The cDNA sequence coding for the chloroplastic [11] and cytosolic [12] isoforms from spinach (*Spinacia oleracea*) have been published. Both the cytosolic and chloroplastic enzymes are encoded by nuclear genes. As with all nuclear-encoded proteins imported into chloroplasts [13], the plastid enzyme is synthesized as a precursor molecule containing a transit peptide [11].

We have now performed amino acid and nucleotide sequence comparisons of the plant and bacterial *O*-acetylserine (thiol)-lyases to characterize the amino acid residues that are important for PLP binding to the plant enzymes and to understand the origin of the chloroplast and cytosolic isoforms. We discuss the results in the frame of the endosymbiotic theory of chloroplast evolution and suggest that a chloroplast transit peptide may have a bacterial origin.

RESULTS

Sequence samples and alignment

We used the Bestfit program [14] to compare the nucleotide and deduced amino acid sequences of the three following *O*-acetylserine (thiol)-lyases: the isoform A from *E. coli* (*cysK* gene) [7], the cytosolic [12] and the chloroplast [11] isoenzymes from spinach leaves. The *O*-acetylserine (thiol)-lyase isoform A from *S. typhimurium* was not considered in this comparison because it exhibits extensive sequence identity (98 %) with isoform A from *Escherichia coli*. A recent report also shows that the amino acid sequence of the enzyme isolated from bell-pepper (*Capsicum annuum*) chromoplasts [15] is almost identical with that of *O*-acetylserine (thiol)-lyase from spinach chloroplasts [11].

When the sequences are aligned, several regions of similarity are found (Figure 1). I corresponds to the entire coding sequence of the bacterial enzyme of which only the 5' end of this region is presented in Figure 1(b). II corresponds to the N-terminal portion of the cytosolic form and includes nucleotides comprised between the initiation codons of the cytosolic and bacterial enzymes. In particular, the AGGA Shine–Dalgarno motif of the bacterial gene is in region II. III and IV contain sequences coding for the N-terminal portion of the chloroplast *O*-acetylserine (thiol)-lyase and for six amino acids at the C-terminal end of the transit peptide of the chloroplast enzyme {the sequence coding for the maturation site is d(AAA↓GCT) [11]}. DNA elements involved in transcriptional control of the bacterial gene (e.g. the '–10' and '–35' boxes) are presumed to be located within region IV [7]. V corresponds to the sequence coding for the major part (46 amino acids out of 52) of the transit peptide of the chloroplast enzyme. A palindromic motif of 22 bases is present in region V of the bacterial sequence, corresponding presumably to a transcription termination site for a gene preceding *cysK* on the *E. coli* genome [7]. Sequence data within V are not yet available for the cytosolic enzyme.

Region I: comparison of coding sequences

Comparison, within entire region I, of the deduced amino acid sequence of the mature form of the amino acid chloroplast *O*-acetylserine (thiol)-lyase revealed an extremely high 73 % nucleic acid sequence identity and 89 % sequence homology, taking into account gaps and equivalent amino acids, with that of the cytosolic enzyme. Analysis of the sequence data of Saito et al. [12] and Rolland et al. [11] also revealed a similar codon bias within region I of the two plant enzymes. In both cases *XYC* and *XYG* are only used 36 % of the time. The low content of C- and G-terminated codons exhibited by the plant enzymes agrees with the codon-choice patterns of spinach proteins as deduced from the GenBank genetic sequence data bank [16]. The amino acid sequence of the two plant enzymes also displays substantial similarity with that of *O*-acetylserine (thiol)-lyase isoform A of *E. coli*. There is 52 % identity and 71 % homology when comparing the bacterial and cytosolic sequences, and 55 % identity and 74 % homology when comparing the bacterial and chloroplast sequences. Altogether comparison of the three sequences discloses 48 % identity and 70 % homology. In contrast with the plant coding sequences, the region predicted to encode the bacterial isoform A shows no codon bias. Here *XYC* and *XYG* are used 51 % of the time.

Pairwise nucleotide sequence identities within the entire region I are as follows: 68 % between the cytosolic and chloroplast sequences, 55 % between the spinach cytosolic and bacterial sequences and 57 % between the spinach chloroplast and the bacterial sequences.

Clearly, from both amino acid and nucleotide sequence analyses, spinach chloroplast *O*-acetylserine (thiol)-lyase is more closely related to the cytosolic enzyme than to the prokaryotic *O*-acetylserine (thiol)-lyase. Although identity between the two plant sequences is extremely high, the fact that they differ by 32 % rules out the possibility that the two plant enzymes are encoded by a single gene. This contention is supported by the results of Southern-blot analyses [11,12].

O-Acetylserine (thiol)-lyases from *E. coli* [17] and spinach leaf chloroplasts [18] have been shown to bind one PLP molecule per enzyme subunit. The PLP-binding-site sequence, as determined by chemical analysis of a pyridoxyl-peptide of the bacterial isoform A, is ILAKVESRNPFSVK⁴³CRIGAN [19]. PLP binds to Lys-43 (i.e. the lysine residue shown in *italic bold*), immediately adjacent to the cysteine residue. We previously noted [11] that the above sequence matches one found around Lys-108 of spinach chloroplast *O*-acetylserine (thiol)-lyase. The chloroplast sequence is IAAKLESMEPCSSVK¹⁰⁸DRIGYS. Such a related sequence (VAAKLEGMEPCSSVK⁴⁹DRIGFS) is also present in spinach cytosolic *O*-acetylserine (thiol)-lyase, located around Lys-49 of this protein. It seems therefore unlikely that the PLP-binding site of spinach cytosolic *O*-acetylserine (thiol)-lyase corresponds to Lys-226, as proposed by Saito et al. [12]. From this comparison it is apparent that the chloroplast enzyme is closely related to the cytosolic enzyme.

Region II: nucleotide sequence coding for the N-terminal portion of spinach cytosolic *O*-acetylserine (thiol)-lyase

Comparison, according to the alignment presented in Figure 1, of spinach cytosolic *O*-acetylserine (thiol)-lyase sequence with that of *E. coli* isoform A revealed the presence of a highly conserved domain, encoding the N-terminal part of the eukaryotic enzyme (region II in Figure 1). The domain, composed of 15 nucleotides, is delimited on the 5' side by the initiation codon of the cytosolic enzyme and on the 3' side by the initiation codon of its bacterial counterpart. Remarkably, among the 15 nucleotides

in the cytosolic sequence, 12 are found in same order in the 5'-untranslated region located immediately upstream from the d(ATG) of the bacterial gene. Introduction of a single gap in the bacterial sequence allows one to reach a very high identity of 80 % (Figure 1b). The consensus sequence of the motif is d(GTTNAGGANNAGGCC). The AGGA Shine-Dalgarno bacterial sequence is entirely conserved in the eukaryotic sequence. The statistical significance of this homology was evaluated by a search of the above motif in the EMBL genetic sequence data bank. Out of 75 000 nucleotide sequences registered in this bank, only the *cysK* gene coding for *E. coli* *O*-acetylserine (thiol)-lyase isoform A [7] exhibited this motif. [Note that the nucleotide sequence encoding spinach cytosolic *O*-acetylserine (thiol)-lyase was not yet present in the EMBL data bank.]

Comparison, within region II of the 5'-untranslated region of the bacterial gene with the nucleotide sequence coding for mature chloroplast *O*-acetylserine (thiol)-lyase revealed a 47 % identity. The best alignment was achieved by introducing several gaps in the chloroplast sequence, resulting mainly in the loss of the AGGA Shine-Dalgarno bacterial sequence. A substantially higher identity of 60 % was obtained by comparing the nucleotide sequences of the spinach cytosolic and chloroplast enzymes (Figure 1b), reinforcing the finding that spinach chloroplast *O*-acetylserine (thiol)-lyase is more closely related to the cytosolic enzyme than to the prokaryotic *O*-acetylserine (thiol)-lyase.

Regions III, IV and V: nucleotide sequence coding for the N-terminal portion of the mature form of spinach chloroplast *O*-acetylserine (thiol)-lyase and for the transit peptide

Spinach chloroplast *O*-acetylserine (thiol)-lyase is synthesized as a precursor protein containing an N-terminal extension of 52 amino acids [11]. This extension exhibits the characteristic features of transit peptides found in nuclear-encoded proteins that are imported into chloroplasts [13]. Upon import into the chloroplasts, the transit peptide is excised, thus freeing the mature enzyme.

To find the possible origin of the N-terminal portion of mature chloroplast *O*-acetylserine (thiol)-lyase and of the transit peptide, we compared the nucleotide sequences according to the alignment shown in Figure 1. Region III, located upstream from region II, contains the initiating codon of the cytosolic enzyme and 17 nucleotides located upstream from this trinucleotide (Figure 1b). In this region there are two potential stop codons within the sequence of the cytosolic enzyme. There is no obvious homology between the three analysed nucleotide sequences: pairwise identities range from 28 to 35 %. Yet, the introduction of several gaps in the chloroplast sequence yielded an upstream motif composed of a stretch of 31 nucleotides, referred to as region IV in Figure 1, where substantially higher homologies were found. Region IV codes for the N-terminal part of the mature chloroplast enzyme and for six amino acids at the C-terminal part of the transit peptide. Here there is 66 % identity between the chloroplast and cytosolic spinach sequences, and 52 % identity between the spinach cytosolic sequence and that of bacterial isoform A. The homology between the spinach chloroplast sequence and that of bacterial isoform A is slightly lower (48 % identity).

Region V that codes for the transit peptide of the chloroplast *O*-acetylserine (thiol)-lyase shows no codon bias: here *XYC* and *XYG* are used 50 % of the time, as found for region I predicted to encode the bacterial isoform A. Comparison of the chloroplast nucleotide sequence with the 5'-non-coding sequence of *O*-acetylserine (thiol)-lyase isoform A from *E. coli* revealed 44 % overall identity. The best alignment was obtained by introducing several gaps in the bacterial sequence. Interestingly, out of the 18

gaps, eight are found clustered within a palindromic sequence of 22 bases, assumed to correspond to a transcription termination site for a gene preceding *cysK* on the *E. coli* genome [7]. Homology between the bacterial and the chloroplast sequences is higher on the right side (50 % identity) than on the left side (39 % identity) of the bacterial palindromic motif. In particular, on the right side of the palindrome there is a stretch of 34 nucleotides where the bacterial and chloroplast sequences disclose 59 % identity.

DISCUSSION

The present study was undertaken to understand better the structure, function and biogenesis of the different forms of *O*-acetylserine (thiol)-lyase found in leaf tissues. The endosymbiotic theory of chloroplast evolution proposes that many nuclear-encoded proteins originated from a prokaryotic ancestor [20,21]. In this context it is noteworthy that, among eukaryotes, only plastid-containing cells are able to synthesize cysteine from *O*-acetyl-L-serine. The precursor of cysteine synthesis in animal and yeast cells is not *O*-acetyl-L-serine but rather methionine, homocysteine or *O*-acetyl-L-homoserine. This is because these cells do not contain any detectable serine transacetylase activity that catalyses the synthesis of *O*-acetyl-L-serine from serine and acetyl-CoA [22].

Sequence data on plant enzymes located in both the cytosol and the chloroplast have shown that chloroplast enzymes may exhibit sequence similarity with either a cytosolic or a bacterial counterpart. Collectively these data have been interpreted according to two possible mechanisms compatible with the endosymbiont hypothesis. The first involves a loss of chloroplast genes and functional replacement by nuclear genes. The second assumes a transfer of chloroplast genes to the nucleus. Support for either one of these two mechanisms was provided by sequence analyses on plant glyceraldehyde-3-phosphate dehydrogenase [21,23], NADP⁺-dependent malic enzyme [24], glutamine synthetase [25], NADP:malate dehydrogenase [26] and pyruvate orthophosphate dikinase [27]. Our results suggest that both mechanisms may have contributed to the evolution of *O*-acetylserine (thiol)-lyase activity in plants. Thus the cDNAs coding for the cytosolic and chloroplastic forms of this enzyme exhibit a high sequence identity (68 %). Also, in contrast with the bacterial gene, both plant genes show the same bias in codon usage. Interestingly, a high identity of 66 % is also found by comparing the sequence coding for the N-terminal part of the mature chloroplast enzyme and the 5'-untranslated region of the cytosolic gene (region IV). These features suggest that the chloroplast *O*-acetylserine (thiol)-lyase originated from the cytosolic *O*-acetylserine (thiol)-lyase. A comparison of the coding regions of the two plant enzymes with that for the bacterial isoform A also reveals a high identity (55–57 %). This identity is not restricted only to the PLP-cofactor-binding domain, but spans the full length of the three analysed sequences. Taken together, these features are compatible with a mechanism in which both plant genes arose from a common prokaryotic ancestor. Furthermore, the present comparison suggests that not only was the coding sequence of the *O*-acetylserine (thiol)-lyase gene of the endosymbiont transferred to the plant cell nucleus, but that this also occurred for part of the 5'-non-coding region of this gene. Thus the N-terminal portion of spinach cytosolic *O*-acetylserine (thiol)-lyase appears to derive in full from the 5'-non-coding region of the corresponding bacterial gene (region II). Also, within the sequence encoding the transit peptide of the chloroplast enzyme, there are stretches of bases showing substantial homology with the 5'-non-coding region of the bacterial

isoform A gene (regions IV and V), suggesting that the transit peptide of chloroplast *O*-acetylserine (thiol)-lyase may have, at least in part, a bacterial origin.

The mechanism involved in the transformation of a non-coding sequence into new coding ones during the course of evolution is presently unknown. However, the phenomenon may not be without precedent. For example, the possibility that generation of new genes may arise by overprinting of pre-existing coding or non-coding nucleotide sequences has already been discussed [28–29]. The fact that the sequence encoding the transit peptide does not show the codon bias observed in the region coding for the mature enzyme may reflect modular evolution for the two functional regions of the spinach chloroplast protein. This would agree with the results of von Heijne et al. [13] showing that transit peptides are remarkably variable in their amino acid sequence, but exhibit common structural domains to optimize the import and maturation processes of chloroplast protein precursors.

Finally, we note that the best alignment of sequences in Figure 1 produced a number of gaps clustered within regions II and III of the chloroplast *O*-acetylserine (thiol)-lyase. Wasmann et al. [30] and Smeekens et al. [31] have shown that, in addition to the transit peptide, the first amino acids contributed by the mature sequences of nuclear-encoded chloroplast proteins, adjacent to the processing site, play an important role in import of such proteins into chloroplasts. These observations have led to the proposal [31] that transit peptides might evolve in such a way as to optimally adapt to (or compensate for) the structure of their respective mature proteins. In this respect one could imagine that the N-terminal part of the mature chloroplast protein has evolved so as to optimize the function of the transit peptide.

We thank Rick DeRose, Bernard Dumas, Renaud Dumas and Michel Lebrun for helpful suggestions and discussions.

REFERENCES

- 1 Becker, M. A. and Tomkins, G. M. (1969) *J. Biol. Chem.* **244**, 6023–6030
- 2 Hulanicka, M. D., Kredich, N. M. and Treiman, D. (1974) *J. Biol. Chem.* **249**, 867–872
- 3 Hulanicka, M. D., Hallquist, S. G., Kredich, N. M. and Mojica-A, T. (1979) *J. Bacteriol.* **140**, 141–146
- 4 Nakamura, T., Iwahashi, H. and Eguchi, Y. (1984) *J. Bacteriol.* **158**, 1122–1127
- 5 Nakamura, T., Kon, Y., Iwahashi, H. and Eguchi, Y. (1983) *J. Bacteriol.* **156**, 656–662
- 6 Byrne, C. R., Monroe, R. S., Ward, K. A. and Kredich, N. M. (1988) *J. Bacteriol.* **170**, 3150–3157
- 7 Levy, S. and Danchin, A. (1988) *Mol. Microbiol.* **2**, 777–783
- 8 Sirko, A., Hryniewicz, M., Hulanicka, D. and Boeck, A. (1990) *J. Bacteriol.* **172**, 3351–3357
- 9 Lunn, J. E., Droux, M., Martin, J. and Douce, R. (1990) *Plant Physiol.* **94**, 1345–1352
- 10 Rolland, N., Droux, M. and Douce, R. (1992) *Plant Physiol.* **98**, 927–935
- 11 Rolland, N., Droux, M., Lebrun, M. and Douce, R. (1993) *Arch. Biochem. Biophys.* **300**, 213–222
- 12 Saito, K., Miura, M., Hirano, H. and Murakoshi, I. (1992) *Proc. Natl. Acad. Sci. U.S.A.* **89**, 8078–8082
- 13 von Heijne, G., Steppuhn, J. and Herrmann, R. G. (1989) *Eur. J. Biochem.* **180**, 535–545
- 14 Smith, T. F. and Waterman, M. S. (1981) *Adv. Appl. Math.* **2**, 482–489
- 15 Römer, S., d'Harlingue, A., Camara, B., Schantz, R. and Kuntz, M. (1992) *J. Biol. Chem.* **267**, 17966–17970
- 16 Wada, K.-N., Wada, Y., Doi, H., Ishibashi, F., Gojobori, T. and Ikemura, T. (1991) *Nucleic Acids Res.* **19** (Suppl.), 1981–1986
- 17 Becker, M. A., Kredich, N. M. and Tomkins, G. M. (1969) *J. Biol. Chem.* **244**, 2418–2427
- 18 Droux, M., Martin, J., Sajus, P. and Douce, R. (1992) *Arch. Biochem. Biophys.* **295**, 379–390
- 19 Nalabolu, S. R., Tai, C.-H., Schnackerz, K. D. and Cook, P. F. (1992) *Amino Acids* **2**, 119–125

-
- 20 Sagan, L. (1967) *Theor. Biol.* **14**, 225–274
- 21 Shih, M. C., Lazar, G. and Goodman, H. M. (1986) *Cell* **47**, 73–80
- 22 Cherest, H. and Surdin-Kerjan, Y. (1992) *Genetics* **130**, 51–58
- 23 Brinkmann, H., Martinez, P., Quigley, F., Martin, W. and Cerff, R. (1987) *J. Mol. Evol.* **26**, 320–328
- 24 Rothermel, B. A. and Nelson, T. (1989) *J. Biol. Chem.* **264**, 19587–19592
- 25 Tingey, S. V., Tsai, F.-Y., Edwards, J. W., Walker, E. L. and Coruzzi, G. M. (1988) *J. Biol. Chem.* **263**, 9651–9657
- 26 Metzler, M. C., Rothermel, B. A. and Nelson, T. (1989) *Plant Mol. Biol.* **12**, 713–722
- 27 Sheen, J. (1991) *Plant Cell* **3**, 225–245
- 28 Ohno, S. (1984) *Proc. Natl. Acad. Sci. U.S.A.* **81**, 2421–2425
- 29 Keese, P. K. and Gibbs, A. (1992) *Proc. Natl. Acad. Sci. U.S.A.* **89**, 9489–9493
- 30 Wasmann, C. C., Reiss, B., Bartlett, S. G. and Bohnert, H. J. (1986) *Mol. Gen. Genet.* **205**, 446–453
- 31 Smeeckens, S., Geerts, D., Bauerle, C. and Weisbeek, P. (1989) *Mol. Gen. Genet.* **216**, 178–182
-

Received 25 January 1993; accepted 8 March 1993